Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Identificação de genes associados com o potencial invasivo de *Streptococcus pneumoniae*

Luísa Moreira Sêco

Dissertação
Mestrado em Bioestatística
2012

Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Identificação de genes associados com o potencial invasivo de *Streptococcus pneumoniae*

Luísa Moreira Sêco

Dissertação orientada pela Prof.ª Doutora Marília Cristina de Sousa Antunes
e pelo Prof. Doutor Francisco Rodrigues Pinto

Mestrado em Bioestatística

2012

Agradecimentos

Agradeço a todas as pessoas que contribuíram, de várias formas, para a realização desta dissertação de Mestrado.

Aos meus orientadores, Professora Doutora Marília Antunes e Professor Doutor Francisco Pinto, pelos conhecimentos que me transmitiram, pela ajuda, e pela disponibilidade que sempre demonstraram ao longo da realização deste trabalho.

Aos docentes do Departamento de Estatística e Investigação Operacional da FCUL, pelos conhecimentos transmitidos ao longo deste Mestrado e pela sua ajuda sempre que foi solicitada.

Aos meus pais e à minha irmã, pelas oportunidades que me proporcionaram e por sempre me terem incentivado a aprender mais.

Aos meus amigos Elsa, Paula, Patrícia e Paulo, pela sua amizade e apoio nos momentos mais difíceis deste percurso.

Agradeço ao Instituto de Microbiologia - Instituto de Medicina Molecular da Faculdade de Medicina da Universidade de Lisboa, pela disponibilização dos dados a partir dos quais foi possível este estudo ser efetuado, e de toda a informação que me foi facultada no decurso deste trabalho.

Resumo

O *Streptococcus pneumoniae* é uma espécie bacteriana que integra a flora comensal da nasofaringe humana e constitui uma causa frequente de doença invasiva nas vias respiratórias. A sua diversidade genética e fenotípica é muito elevada, conhecendo-se estirpes invasivas (patogénicas) e estirpes colonizadoras (assintomáticas). A rápida evolução do genoma desta espécie originada por transformação genética, e o consequente aparecimento de novas estirpes conduz à necessidade de identificar os genes que estão associados à sua capacidade invasiva.

Esta dissertação tem como objetivo identificar genes associados ao potencial invasivo de *Streptococcus pneumoniae* a partir de dados de hibridação genómica comparativa em que foi pesquisada a presença de 3620 genes em 72 estirpes, classificadas como invasivas, colonizadoras ou neutras com base em estudos epidemiológicos. A pesquisa de genes associados à invasibilidade foi efetuada através de classificação não supervisionada através de agrupamento hierárquico, e classificação supervisionada através da construção de modelos em árvore.

A classificação não supervisionada das estirpes com base no seu conteúdo genético demonstrou não ser um método eficaz para detetar genes associados ao seu potencial invasivo.

Os resultados da classificação supervisionada revelaram que a construção de modelos de classificação em árvore, quando efetuada a partir de variáveis (genes) previamente selecionadas, constitui a forma mais eficaz de encontrar associações entre a presença/ausência de determinados genes e o potencial invasivo das estirpes.

Os modelos em árvore que demonstraram uma melhor capacidade preditiva e maior robustez resultaram da classificação a partir de uma matriz indicadora da presença dos genes. Concluiu-se também que não é necessário um número elevado de variáveis para obter modelos com a máxima capacidade preditiva.

Palavras-chave: *Streptococcus pneumoniae*, *microarrays*, hibridação genómica comparativa, agrupamento hierárquico, modelo em árvore.

Abstract

Streptococcus pneumoniae is a bacteria species in the commensal flora of the human nasopharynx, thus being a frequent cause of invasive diseases in the respiratory tract. This species of bacteria has a large genetic and phenotypic diversity, where invasive (pathogenic) and colonizer (asymptomatic) strains are known. The rapid evolution of the genome of this species, due to genetic transformation, and the consequent appearance of new strains lead to the need to identify the genes associated to its invasiveness.

This dissertation aims to identify genes associated to the invasive potential of *Streptococcus pneumoniae*, based on comparative genomic hybridization data, where the presence of 3620 genes in 72 strains has been examined. Based upon epidemiological studies, the strains have been classified as invasive, colonizer or neutral. The search for genes associated to invasiveness has been done by means of unsupervised classification through hierarchical clustering and by supervised classification through tree-based models.

The unsupervised classification of strains based on their genetic content has proven not to be an effective method to detect genes associated with their invasive potential.

The results of supervised classification have shown that the tree-based model technique, when based upon previously selected variables (genes), constitutes the most effective way to find associations between the presence / absence of certain genes and the invasive potential of the strains.

The tree-based models which have shown a higher predictive power and a stronger robustness, resulted from the classification from an indicative matrix of the presence of genes. It has also been concluded that there is no need for a large number of variables in order to to obtain models with the highest predictive power.

Keywords: *Streptococcus pneumoniae*, microarrays, comparative genomic hybridization, hierarchical clustering, tree-based model.

Glossário

Ácido desoxirribonucleico (DNA) – Molécula em forma dupla hélice que constitui a parte fundamental dos genes e é responsável pela informação hereditária. É formada por **nucleótidos**, que são moléculas constituídas por um grupo fosfato, uma pentose (desoxirribose) e uma base, que pode ser adenina, timina, citosina ou guanina. A dupla hélice é formada por duas cadeias complementares de nucleótidos, em que a timina é complementar à adenina e a citosina à guanina, encontrando-se ligadas por pontes de hidrogénio ao longo da cadeia.

Ácido ribonucleico (RNA) – Ácido nucleico de cadeia simples de constituição semelhante ao DNA, com a diferença que a pentose é uma ribose e contém a base uracilo em vez da timina.

Aminoácido – Moléculas que constituem a unidade estrutural das proteínas. Contêm um átomo de carbono central ao qual se encontram ligados um grupo amina e um grupo carboxilo.

Codão – Pequena sequência de DNA constituída por três pares de bases que constitui um código para um aminoácido. Durante a síntese proteica os aminoácidos são adicionados à cadeia polipeptídica de acordo com a sequência de nucleótidos do RNA mensageiro.

Cromossoma - Agrupamento linear de genes e outras sequências de DNA, por vezes associado a proteínas e RNA.

Electroforese – Técnica em que são separados os componentes de uma mistura de moléculas (RNA, DNA ou proteínas) através de um campo elétrico aplicado num gel. A mistura de DNA, RNA ou proteínas é colocada no gel, através do qual as moléculas migram e se separam de acordo com as suas diferenças de peso molecular e carga elétrica.

Hibridação – processo através do qual uma cadeia de ácido nucleico se liga a outra cadeia cuja sequência de nucleótidos é complementar à primeira, através do estabelecimento de pontes de hidrogénio entre os nucleótidos que se complementam (adenina com timina e citosina com guanina, no caso do DNA). Este princípio é aplicado aos *microarrays*. A hibridação de uma amostra de DNA com os oligonucleótidos presentes num *spot* só é possível se as suas sequências forem complementares, ou seja, se o gene que o oligonucleótido representa estiver presente na amostra.

Hibridação Genómica Comparativa (*Comparative Genomic Hybridization* ou CGH) - Técnica inicialmente desenvolvida com a finalidade de analisar variações no número de cópias do DNA em cromossomas de células tumorais. O princípio da técnica é a utilização de DNA controlo (amostra de referência, neste caso proveniente de células saudáveis) e DNA da amostra teste, que são marcados com moléculas fluorescentes diferentes (quando excitadas emitem fluorescência em comprimentos de onda diferentes). A técnica começou por ser utilizada utilizando como DNA alvo cromossomas em metafase, aos quais o DNA das células tumorais hibrida competitivamente com o DNA controlo (nas zonas em que são complementares), permitindo identificar alterações no número de cópias das células tumorais através das diferenças das intensidades das fluorescências entre o DNA destas e da amostra de referência. Mais tarde foi desenvolvida a CGH em *microarrays* em que, em vez de utilizar cromossomas em metafase como alvo, utiliza *slides* (suportes geralmente de vidro) aos quais se encontram fixados oligonucleótidos, que constituem o DNA alvo.

Oligonucleótidos – Pequena sequência de DNA sintético.

***Open reading frame* (ORF)** – Porção de DNA que, quando transcrita em aminoácidos, não contém nenhum codão *stop* (sequência de três nucleótidos que constitui um código para finalizar a transcrição), dando origem a uma cadeia polipeptidica. Uma ORF longa poderá corresponder a parte de um gene.

Pixel – O menor elemento que forma uma imagem digital.

Polipéptido – Cadeia de aminoácidos ligados; pode ser uma proteína ou parte de uma proteína.

Proteína – Molécula constituída por aminoácidos. Pode ter uma função estrutural, contribuindo para as propriedades físicas das células ou organismos, ou funcional, quando desempenha um papel nas reações químicas que ocorrem nas células.

RNA mensageiro (**mRNA**)– Molécula de RNA transcrita a partir do DNA de um gene, a partir da qual é sintetizada uma proteína.

Transcrição – Síntese de RNA a partir de uma cadeia de DNA molde.

Conteúdo

Lista de Figuras

Lista de Tabelas

1. Introdução

1.1. Contextualização do estudo

1.1.1. *Streptococcus pneumoniae*

Streptococcus pneumoniae, conhecido como pneumococo, é uma espécie bacteriana que integra a flora comensal da nasofaringe humana, e constitui uma causa frequente de doença invasiva nas vias respiratórias, podendo ser responsável por infecções graves - meningite, sépsis ou pneumonia – e outras infecções menos graves, como sinusite e infecções do ouvido médio. Estas infecções atingem com maior frequência crianças e idosos, e observam-se em países com diferentes índice de desenvolvimento.

A diversidade genética e fenotípica desta espécie é muito elevada, conhecendo-se estirpes invasivas (patogénicas), e estirpes colonizadoras que fazem parte da flora comensal da nasofaringe mas não originam infecções, sendo a sua presença assintomática.

A frequência da colonização da nasofaringe pelo *Streptococcus pneumoniae* apresenta diferenças etárias, sendo maior nas crianças, com um máximo de 55% aos 3 anos, diminuindo gradualmente até 8% aos 10 anos, idade em que estabiliza (Bogaert et al., 2004).

1.1.2 Variabilidade genética do *Streptococcus pneumoniae*

Durante o último século diversas estirpes isoladas de *Streptococcus pneumoniae* foram categorizadas, por métodos serológicos, tendo sido descritos pelo menos 93 serotipos (Sá-Leão et al., 2011; Calix, 2010) identificados com base nas características químicas e imunológicas da cápsula polissacarídea que envolve a bactéria e a protege da fagocitose. As vacinas anti-pneumocócicas existentes no mercado são preparadas de acordo com o conhecimento prévio acerca da invasibilidade e frequência dos diferentes serotipos nas populações. No entanto, existem estirpes que não possuem cápsula polissacarídea, não estando assim incluídas nesta classificação. Estas estirpes não são virulentas uma vez que não

estão protegidas da fagocitose no hospedeiro; um exemplo é a estirpe R6.

A variabilidade observada dentro desta espécie, e as consequentes adaptações às condições ambientais, têm origem nas mutações originadas por substituição de nucleótidos, rearranjos no DNA e aquisição de genes. Os dois últimos fenómenos podem resultar de transformação genética, em que podem ser introduzidas, através de recombinação, sequências de DNA que eram funcionais no dador. A substituição simultânea de componentes de complexos proteicos (mesmo quando codificados por genes não ligados ao cromossoma) pode ocorrer durante o processo de transformação. A incorporação de DNA a partir de outros microrganismos da mesma espécie ou mesmo de espécies diferentes do mesmo nicho ecológico constitui assim um mecanismo poderoso de rápida evolução do genoma, e pode contribuir de forma significativa para a fluidez genómica no *Streptococcus pneumoniae*. (Claverys 2000). Segundo Hiller (2010), o aparecimento de novas estirpes bacterianas deve-se à ocorrência de transferência horizontal de genes durante o processo infecioso.

É possível caracterizar as estirpes através da classificação em clones, baseada em estudos de PulseFieldGelElectrophoresis, que consiste na restrição enzimática do DNA e subsequente electroforese. O padrão de bandas observado na electroforese irá apresentar diferenças, e esse padrão poderá ser utilizado para caracterizar diferentes clones, sendo um clone um conjunto de estirpes com um padrão de bandas semelhante.

1.1.3. Hibridação Genómica Comparativa em *microarrays*

Atualmente existem estirpes de *Streptococcus pneumoniae* totalmente sequenciadas, como é o caso das estirpes R6, G54 e TIGR4, mas a sequenciação de todas as estirpes conhecidas e o estudo de todos os seus genomas é um processo demasiado dispendioso e demorado.

A técnica de **hibridação genómica comparativa em *microarrays*** (*comparative genomic hybridization* ou CGH), permite ultrapassar essa dificuldade, uma vez que torna possível a obtenção de informação acerca da composição genética de um grande número de estirpes da espécie em estudo, que por sua vez permite estimar as probabilidades de milhares de genes estarem presentes ou ausentes em cada estirpe. Esta técnica é também designada por genomotipagem, e é largamente utilizada em estudos do genoma de espécies bacterianas há cerca de uma década.

A tecnologia de *microarrays* foi criada inicialmente com a finalidade de estudar a expressão génica com base na concentração de mRNA (RNA mensageiro) em células tumorais. Nestes casos a expressão de alguns genes encontra-se alterada, e a concentração de cDNA[1] detetada representa o nível a que os genes se encontram ativos, em transcrição. A expressão génica anómala é uma medida de disfuncionalidade das células em estudo.

Um *microarray* é, em termos funcionais, um suporte de vidro, de tamanho e forma semelhante a uma lâmina de microscópio, no qual se encontram fixados individualmente milhares de oligonucleótidos diferentes, ou sondas, (porções de sequências de DNA complementares às sequências dos genes que se pretendem detetar), colocados em pontos de dimensão microscópica ao longo da lâmina designados por *spots*, sendo cada *spot* constituído por milhares de cópias de uma sonda oligonucleotídica. As sondas são fixadas à superfície de vidro através de impressão robótica.

No caso dos estudos de expressão génica, ao incubar o *microarray* com cDNA das células em estudo, os fragmentos de DNA obtidos a partir das células complementares aos oligonucleótidos dos *spots* irão hibridar. A medida em que esta ligação ocorre nos diferentes *spots* pode ser obtida mediante a marcação do cDNA da amostra em estudo com moléculas fluorescentes, e a fluorescência detetada, após lavagens, irá ser proporcional ao DNA da amostra que se ligou. Assim, uma fluorescência elevada indica uma concentração elevada de DNA. A presença de DNA controlo proveniente de células saudáveis, marcado com moléculas de fluorescência diferente permite a obtenção de níveis de fluorescência correspondentes à expressão génica não alterada. É a partir do rácio das intensidades das fluorescências do controlo e do teste que é estimada a probabilidade de cada um dos genes estar a ser diferencialmente expresso nas células tumorais.

No caso da aplicação da tecnologia CGH à composição genética, o princípio e o procedimento são semelhantes, mas neste caso as diferenças de fluorescência representam diferenças na composição génica da amostra e do controlo (se um gene estiver ausente não ocorre hibridação no respetivo *spot*, e a fluorescência da amostra detetada nesse *spot* irá ser reduzida ou ausente).

O *microarray* necessita, por isso, de equipamento robótico especializado, de equipamento de

1 cDNA - molécula de DNA sintetizada a partir de uma molécula de RNA mensageiro que lhe serve de molde (mRNA) através da enzima transcritase reversa. Por essa razão, é designado por DNA complementar ou cDNA.

leitura (*scanner*) e de ferramentas informáticas que permitam registar os resultados obtidos e fazer os cálculos necessários antes da análise de dados.

Na hibridação genómica comparativa de *microarrays* são utilizadas duas amostras de DNA – a amostra de **controlo** e a amostra de **teste**. O genoma da amostra controlo é conhecido, e as sequências dos oligonucleótidos presentes no *microarray* são complementares aos genes que se pretende estudar; logo o *microarray* deve conter oligonucleótidos de genes presentes na amostra controlo, para que haja hibridação em todos os *spots* e leitura de fluorescência. O genoma da amostra teste não é conhecido, pelo que os genes em estudo poderão fazer ou não parte da mesma. Caso um gene não se encontre presente, não ocorrerá hibridação no respetivo *spot*, não sendo produzida fluorescência resultante da amostra de teste. Para que seja possível distinguir a amostra do controlo, as respetivas moléculas de DNA são marcadas com **fluorocromos** diferentes, que são moléculas que emitem fluorescência quando excitadas por raios *laser*. Os fluorocromos utilizados para o efeito são Cy5, que emite fluorescência com comprimento de onda de 670 nm (vermelho) quando excitado por raios *laser*, e Cy3, que emite fluorescência com comprimento de onda de 570 nm (verde).

Durante a incubação do *microarray* com a amostra e o controlo, o DNA marcado de ambos ligar-se-á às sondas oligonucleotídicas complementares nos *spots*. No final do processo o *microarray* é sujeito a lavagens para ser retirado o excesso de DNA não ligado, e colocado num *scanner* que, através de um laser, excita os fluorocromos ligados aos *spots*, que emitem fluorescência no comprimento de onda correspondente, cuja intensidade é lida e registada pelo *scanner*.

Se, por exemplo, a amostra controlo tiver sido marcada com Cy3 e a amostra teste com Cy5, os *spots* correspondentes a genes ausentes na amostra apresentarão coloração verde, uma vez que apenas o DNA do controlo hibridou com os oligonucleótidos. Se a amostra teste contém o referido gene, o DNA do controlo e do teste irão competir pela hibridação com os oligonucleótidos do *spot*. No caso de o DNA da amostra e do teste se encontrarem em iguais proporções, a fluorescência resultante apresentará uma coloração amarela, resultante da mistura da emissão das duas fluorescências. Se a amostra de teste estiver em maior proporção, ligar-se-á também em maior proporção aos oligonucleótidos do *spot*, que apresentará uma coloração mais próxima do vermelho.

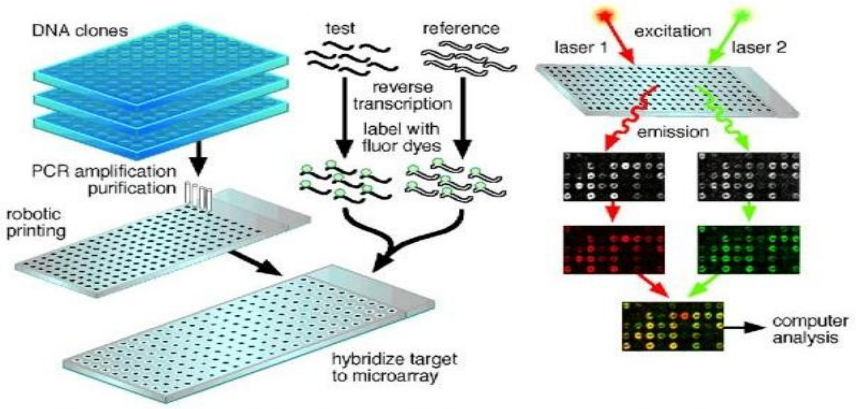


Figura 1.1: Construção de um microarray de dois canais. (Imagem do web site *de BME 240 - Introduction to Clinical Medicine – Universidade da Califórnia).* *http://bme240.eng.uci.edu/students/08s/jentel/Diagnose-hereditary-disease.htm*

O rácio das fluorescências, ou dos seus logaritmos de base 2[2], correspondentes ao teste e ao controlo em cada *spot*, permitirão estimar a probabilidade de o respetivo gene se encontrar presente ou ausente na amostra. No estudo desenvolvido por Cardoso (2009) foram testadas as seguintes medidas:

- $R = \frac{intensidade\ do\ teste}{intensidade\ do\ controlo}$
- $RI = \frac{intensidade\ do\ controlo}{intensidade\ do\ teste}$
- $LR = \log_2\left(\frac{intensidade\ do\ teste}{intensidade\ do\ controlo}\right)$
- $LRI = \log_2\left(\frac{intensidade\ do\ controlo}{intensidade\ do\ teste}\right) + c$, em que c é uma constante tal que $LRI>0$ para cada gene.

O rácio *RI* demonstrou ser a medida que dá origem a resultados em que melhor são discriminados os genes ausentes dos genes presentes, e por essa razão foi a medida utilizada no presente estudo.

2A transformação logarítmica permite transformar dados que muitas vezes são enviesados à direita em dados com distribuição simétrica, de forma a satisfazer o requisito de normalidade. A utilização da base 2 foi escolhida em estudos de expressão génica diferencial, de forma a que a sub-expressão ou sobre-expressão génica fossem comparáveis de uma forma linear. Por exemplo, se R=2, log_2R=1, o que corresponde a uma expressão génica aumentada para o dobro relativamente ao controlo. Se R=0.5 (expressão génica reduzida para metade), log_2R=-1.

1.2. Descrição do Problema

O aparecimento de doença invasiva no Homem após a colonização por *Streptococcus pneumoniae* depende de vários fatores, alguns inerentes ao próprio hospedeiro (sistema imunitário, composição da flora colonizadora do aparelho respiratório superior), mas sobretudo da capacidade invasiva da estirpe em questão. O conhecimento do serotipo da cápsula é uma informação importante, uma vez que a doença invasiva é provocada, na grande maioria dos casos, por estirpes pertencentes a um conjunto de serotipos relativamente pequeno (Obert, 2006). No entanto, cada serotipo comporta uma grande variedade de estirpes diferentes, sendo algumas invasivas e outras não invasivas, e por esse motivo o seu conhecimento por si só pode ser insuficiente para prever o desfecho de uma colonização por *Streptococcus pneumoniae*. A ocorrência frequente de transferência horizontal de genes nesta espécie aumenta a probabilidade de aquisição, por parte de qualquer estirpe, de genes cuja expressão contribua para uma maior capacidade invasiva do microrganismo. Perante este facto, torna-se fundamental identificar genes ou conjuntos de genes cuja presença ou ausência esteja associada à ocorrência de doença invasiva.

O presente trabalho foi efetuado com base em dados de hibridação genómica comparativa em *microarrays*, em que estes foram preparados com oligonucleótidos correspondentes a 3620 genes do *Streptococcus pneumoniae* e hibridados com DNA proveniente de 72 estirpes. Estão também disponíveis dados epidemiológicos das estirpes em estudo (razão das chances, ou *OR*). As estirpes encontram-se classificadas em três categorias definidas pelos valores de *OR* obtidos: **colonizadoras**, **invasivas**, ou **neutras** (Sá-Leão et al., 2011). Conhecida assim a capacidade invasiva de várias estirpes através dos dados epidemiológicos e o genoma dessas estirpes através da genomotipagem, é possível estudar as possíveis associações que existam entre a presença/ausência de genes e a probabilidade de ocorrência de doença invasiva.

1.3 Metodologias adotadas na análise de dados de Hibridação Genómica Comparativa de microarrays

Existem diversas metodologias que podem ser utilizadas de forma a classificar os genes como presentes ou ausentes. Algumas das mais utilizadas e que demonstram melhor capacidade de discriminação são as que se baseiam em modelos de mistura (Newton et al, 2001; Efron et al., 2003; Broët et al., 2002; Lee et al., 2002, Lönnstedt e Speed, 2002; Pan et al, 2002; Kendziorski et al., 2003; Storey e Tibshirani, 2003; Dean e Raftery, 2005; Antunes e Sousa, 2008). Estas metodologias foram inicialmente desenvolvidas para o estudo da expressão génica diferencial, demonstrando igualmente bons resultados na classificação de genes em ausentes/presentes em estudos de genomotipagem (Cardoso, 2009).

Partindo do pressuposto que existem duas subpopulações (genes com expressão normal/diferencial ou genes ausentes/presentes), é possível encontrar um nível de fluorescência que marca a "fronteira" entre essas subpopulações. Através de métodos iterativos é determinado esse ponto de corte, bem como a probabilidade de cada gene se encontrar diferencialmente expresso ou ausente, em estudos de expressão génica ou de classificação de genes, respetivamente.

Para que se obtenha resultados que permitam classificar com a máxima precisão cada um dos genes em ausentes ou presentes, é necessário dispôr de um controlo que permita obter resultados com um nível de precisão elevado para o maior número de genes possível. No presente estudo foi utilizado um controlo obtido a partir de uma mistura equimolar de DNA das estirpes TIGR, R6 e G54, com base nos resultados de Pinto et al. (2008). Este controlo misto traz vantagem sobre um controlo obtido apenas a partir de uma estirpe, uma vez que conduz à diminuição na taxa de erro nos genes que fazem parte do genoma essencial[3] da espécie: o conjunto de genes presentes nas três estirpes permite obter um controlo mais abrangente, pois pelo menos uma das estirpes contém cada um dos genes. O ganho em sensibilidade obtido pelo controlo misto em relação ao controlo apenas com a estirpe TIGR4 (aumento em 21%, p=0,001) compensa a perda em especificidade (diminuição em 5%, p=0,014) (Pinto et. al. ,2008).

No processo de otimização da metodologia utilizada no presente trabalho (Cardoso, 2009) foi

3 Genoma essencial – Tradução de *core genome*, que designa o conjunto dos genes que são essenciais ao microrganismo e, por essa razão, se encontram presentes em todas as estirpes.

testada a capacidade discriminativa de três modelos de mistura:

- Modelo de Mistura Normal-Uniforme (Dean e Raftery, 2005) – Parte do pressuposto que a população de genes da estirpe em estudo é constituída por uma supopulação de genes presentes, que apresenta distribuição Normal, e uma subpopulação de genes ausentes, com distribuição Uniforme. Através da aplicação do algoritmo de Estimação-Maximização (EM) são estimados os parâmetros das duas subpopulações e atribuídas, aos genes, as respetivas probabilidades a *posteriori* de pertencerem à subpopulação Uniforme.

- Modelo de Mistura Gama-Gama (Newton et al., 2001; Kendziorski et al., 2003; Antunes e Sousa, 2008) – Este método parte do pressuposto que as duas subpopulações em estudo podem ser modeladas através de duas distribuições Gama. Os parâmetros das duas subpopulações e as probabilidades a *posteriori* de cada gene pertencer à subpopulação dos genes ausentes são estimadas através do algoritmo EM.

- Classificador Bayesiano (Antunes e Sousa, 2008) – Ao contrário dos métodos anteriores, que não carecem de informação acerca dos dados - classificação não supervisionada -, neste método a classificação é feita com base numa amostra para a qual a classificação dos genes é conhecida, denominada conjunto de treino – classificação supervisionada (Dudoit e Fridlyand, 2003). Esta metodologia pressupõe a existência de uma característica contínua (X) que é mensurável na população em estudo. A população de genes é dividida em j grupos (neste caso j=2), em que X se comporta de forma diferente consoante pertence a cada um dos grupos. Neste caso X é uma medida que irá permitir estimar a probabilidade de ausência dos genes. O conjunto de treino L é um grupo de genes para os quais a classificação é já conhecida, assumindo os valores de 0 (genes asusentes) ou 1 (genes presentes). Através do conjunto de treino são calculadas as probabilidades preditivas condicionais, e a distribuição preditiva obtida serve posteriormente para calcular a regra de classificação que vai ser aplicada aos restantes genes da amostra (Cardoso, 2009).

No estudo desenvolvido por Cardoso L. (2009), os modelos de mistura NU e GG, bem como os mesmos modelos em conjunto com o classificador bayesiano, revelaram uma exatidão semelhante no que diz respeito em termos de capacidade discriminativa (87 % no caso dos modelos de mistura com ou sem a utilização do classificador bayesiano, e 86% usando apenas o classificador bayesiano).

Com base nestes resultados é possível, aplicando um dos modelos de mistura aos dados (Gama-Gama ou Normal-Uniforme), obter os resultados das probabilidades de ausência dos

genes com a melhor capacidade discriminativa possível, antes de prosseguir para a análise dos dados.

1.4 Objetivos

O objetivo primordial do presente estudo é, com base no conjunto de dados de hibridação genómica comparativa de *microarrays* nos quais são analisados 3620 genes em 72 estirpes, identificar os genes que se encontram associados à capacidade invasiva do *Streptococcus pneumoniae*. As 72 estirpes em estudo fazem parte de uma coleção que abrange estirpes invasivas e colonizadoras assintomáticas, de modo a ter representantes de linhagens genéticas predominantemente invasivas, predominantemente colonizadoras ou intermédias (neutras) entre estes dois extremos.

Pretende-se, assim, averiguar uma possível associação entre a natureza das estirpes no que diz respeito à sua classificação epidemiológica e o seu conteúdo genético e, com base nessa associação, construir um modelo de classificação que permita prever, com um nível de precisão suficientemente informativo, a capacidade invasiva de qualquer estirpe com base no seu perfil genético. No entanto, entre os 3620 genes do *microarray*, que constituem as variáveis do estudo, alguns serão mais informativos que outros, na medida em que não se encontram presentes em todas as estirpes, sendo assim responsáveis pelas diferenças intra-espécie. Nas espécies bacterianas existem genes que se encontram em todas as estirpes, por serem essenciais às funções celulares ou codificarem proteínas estruturais essenciais, e genes que poderão não estar presentes em todas as estirpes, sendo, por isso, responsáveis pelas diferenças intra-espécie. O primeiro grupo de genes é designado por ***genoma essencial***, e o segundo por ***genoma acessório***. É neste último que deverão ser procurados aqueles que podem estar associados à capacidade invasiva desta espécie.

O objetivo inicial foi, pois, identificar o genoma acessório de forma a reduzir o número de variáveis excluindo as que são pouco informativas. De seguida, a partir dos respetivos dados, procurar possíveis associações entre a presença e ausência dos genes e a classificação epidemiológica das estirpes, e construir um modelo que permita classificar qualquer estirpe de *Streptococcus pneumoniae* quanto à sua capacidade invasiva, com base na sua composição genética. Há que salientar que quanto menor for o número de genes cuja presença é

necessário conhecer numa estirpe a classificar, mais vantajoso é o modelo, pois é mais acessível e menos dispendioso laboratorialmente pesquisar a presença de um pequeno número de genes do que um número de genes elevado. Esta informação terá, assim, uma importância clínica relevante ao permitir estimar a capacidade invasiva de uma estirpe desconhecida.

1.5. Metodologias e procedimentos adotados ao longo do estudo

A partir dos resultados de hibridação genómica de *microarrays* foi obtida uma matriz com as intensidades médias de fluorescência referentes a todos os oligonucleótidos. As probabilidades de cada gene se encontrar ausente em cada uma das estirpes foram calculadas a partir do algoritmo EM com base no modelo de mistura Normal-Uniforme. Neste caso a componente Normal da população corresponde aos genes presentes e a componente Uniforme aos genes ausentes.

Estimadas assim as probabilidades de ausência de cada um dos genes em cada uma das estirpes, obteve-se uma matriz 72×3620, em que as 72 estirpes constituem os objetos do estudo, e os 3620 genes as variáveis. Todo o estudo foi efetuado a partir desta matriz de dados.

Sabendo que nem todas as 3620 variáveis são informativas, o primeiro procedimento foi reduzir a matriz inicial às variáveis (genes) que poderão estar associadas ao potencial invasivo das estirpes. Assim, com base no pressuposto da existência do genoma essencial e do genoma acessório, a matriz de dados foi reduzida aos genes potencialmente pertencentes ao genoma acessório, através da eliminação dos genes que apresentavam uma probabilidade inferior a 0.5 de se encontrarem ausentes em todas as estirpes. Assim, a matriz resultante, a partir da qual se trabalhou, ficou reduzida a 1775 variáveis (genes).

Após a redução dos dados ao genoma acessório, foram utilizadas e testadas várias metodologias de classificação das estirpes. Numa primeira primeira fase foi efetuada a **classificação não supervisionada** das estirpes com base na sua composição genética, utilizando métodos de classificação hierárquica, para comparar os grupos obtidos com a classificação das estirpes com base no seu potencial invasivo. Os resultados mostram que os

grupos resultantes da classificação hierárquica não coincidem com a classificação das estirpes, tendo-se prosseguido para a **classificação supervisionada** das estirpes através da elaboração de modelos em árvore.

Os modelos em árvore são construídos através de particionamento de forma recursiva dos objetos em estudo (estirpes) com base nas variáveis que, sucessivamente, apresentam uma maior associação com o desfecho de cada um dos objetos. A seleção da variável (gene) que determina o particionamento em cada fase do algoritmo é escolhida com base numa medida de impureza cujo valor é tanto maior quanto maior a associação entre a variável e o desfecho de cada um dos objetos – **entropia**. Esta metodologia tem a vantagem de selecionar, nos vários passos do algoritmo, a variável mais informativa, o que resulta na redução significativa do número de variáveis necessárias para obter uma classificação. Através do *software* R e da biblioteca RPART [4] foram construídos vários modelos de classificação e testados através de validação cruzada de forma a encontrar aquele que produz uma melhor precisão e melhores valores preditivos ao classificar uma possível estirpe desconhecida como invasiva, colonizadora ou neutra.

Os métodos de classificação com base em análise de dados multivariados são com frequência utilizados no estudo de dados de CGH, nomeadamente a Análise de Componentes Principais (ACP), pois permite reduzir a dimensionalidade dos dados e identificar as variáveis mais informativas. Assim, foi ainda aplicado o modelo de classificação em árvore obtido neste estudo aos dados transformados através da ACP, de forma a comparar o seu desempenho com o modelo aplicado aos dados não transformados.

No capítulo 2 (2.1 e 2.2) é apresentada a descrição dos dados e da forma como estes foram obtidos (médias de *RI*, probabilidades de ausência e classificação das estirpes) e a respetiva análise exploratória. Em 2.3 é feito um resumo teórico dos métodos de classificação não supervisionada (algoritmos baseados em particionamento) e classificação supervisionada (modelos em árvore) e das suas aplicações.

Na secção 2.3.3 são descritos os algoritmos utilizados na construção dos modelos em árvore, utilizando todas as variáveis ou variáveis pré-selecionadas.

No capítulo 3 são apresentados os resultados da classificação não supervisionada das estirpes (3.1), seguidos dos resultados da classificação supervisionada através de modelos em árvore

4RPART (*recursive partioning*) - Biblioteca criada para a construção de árvores de classificação e regressão, na qual foi implementado o algoritmo CART (Classification and Regression Trees, Breiman et. al, 1984).

(3.2) onde são apresentados e comparados os resultados da validação cruzada dos vários modelos através do cálculo da sua precisão e dos valores preditivos para as estirpes colonizadoras, invasivas e neutras (*VPC*, *VPI* e *VPN*, respetivamente). Por fim é feita a discussão dos resultados em que é comparado o desempenho dos diferentes métodos (3.3). No capítulo 4 são apresentadas as conclusões e algumas considerações finais.

2. Materiais e Métodos

2.1. Trabalho laboratorial e tratamento prévio dos resultados

Esta seção descreve sucintamente o trabalho que deu origem aos dados utilizados neste estudo, que foi realizado no Instituto de Microbiologia - Instituto de Medicina Molecular da Faculdade de Medicina da Universidade de Lisboa, em particular pelos investigadores Sandra Aguiar, Professor Mário Ramirez e Professor José Melo-Cristino.

Nas experiências de CGH foram utilizados *microarrays* com 16228 *spots* correspondentes a quatro réplicas de 4057 oligonucleótidos com 70 pb (pares de bases), cada um deles complementar de uma *open reading frame* (ORF) do genoma do *Streptococcus pneumoniae*.

As amostras de controlo utilizadas nas experiências foram obtidas a partir de uma mistura equimolar de DNA proveniente das estirpes TIGR, R6 e G54, de acordo com os estudos prévios (Pinto et al., 2008 e Cardoso L., 2009).
O DNA das amostras de teste foi extraído a partir de 72 estirpes isoladas em laboratório, provenientes de colheitas nasofaríngeas. As 72 estirpes são representantes das diversas linhagens genéticas de *Streptococcus pneumoniae* encontradas na população portuguesa (Sá-Leão et al., 2011).
No total foram feitas 72 experiências, cada uma delas correspondente a uma estirpe em estudo. A escolha dos fluorocromos para o controlo e o teste foi aleatória em cada uma das experiências, de forma a compensar o efeito das diferenças de intensidade da fluorescência associadas aos dois fluorocromos, Cy3 e Cy5. As experiências foram efetuadas com o equipamento Agilent Technologies (câmara de hibridação e *scanner*). As imagens foram analisadas através do *software* Feature Extraction 9.1 (Agilent Technologies). O sinal de cada *spot* foi corrigido através do método de *background correction*, no qual as intensidades médias de cada *spot* correspondentes a cada um dos canais (*red* e *green*) são subtraídas pela média das intensidades lidas fora da área dos *spots* (*background*). As médias são calculadas a partir das intensidades lidas em cada um dos pixeis. Esta correção destina-se a eliminar o efeito da fluorescência que não resultou da ligação do DNA aos oligonucleótidos do *microarray* (originada, por exemplo, por lavagem insuficiente ou partículas de pó).
Depois de obtidas por este processo as intensidades corrigidas, foi-lhes aplicado um modelo

espacial para corrigir efeitos espaciais que possam existir. O enviesamento resultante das intensidades da fluorescência foi de seguida removido através de regressão *loess*[5], obtendo-se assim as intensidades normalizadas, a partir das quais se prosseguiu com a análise.

Os resultados do *software* Feature Extraction são apresentados na forma de uma tabela com os valores das intensidades médias, intensidades medianas e intensidades normalizadas lidas em cada um dos canais (*red* e *green*, correspondentes a Cy5 e Cy3). A partir das intensidades normalizadas foram calculados os rácios de intensidade *RI* para cada *spot* i e calculadas as médias de *RI* para cada uma das quatro réplicas dos oligonucleótidos do *microarray*, de acordo com o procedimento adotado em Cardoso (2009). Este procedimento foi efetuado para cada uma das 72 experiências correspondentes às 72 estirpes.

A classificação dos genes foi efetuada através da aplicação do modelo de mistura Normal-Uniforme às médias das razões da intensidade através do algoritmo de Estimação-Maximização (EM), que permite calcular as probabilidades de ausência de cada gene em cada uma das estirpes.

2.2. Descrição detalhada dos dados

Os dados fornecidos para o presente estudo resultaram das 72 experiências de CGH, e encontram-se reunidos numa matriz composta pelas médias das razões das intensidades da fluorescência (*RI* média). Foi também fornecida a matriz das probabilidades de ausência de cada gene em cada estirpe, e uma matriz de dados biológicos e epidemiológicos referentes a cada uma das estirpes: serotipo da cápsula, clone, razão das chances (*odds ratio* ou *OR*) e classificação (invasiva, colonizadora ou neutra).

Na tabela 2.1 é apresentada uma parte dos dados da matriz das médias de *RI* respeitantes a cada oligonucleótido.

5 *Loess* (*local weighted polynomial regression)* – Regressão não linear em que é ajustada uma função polinomial aos dados, ponderada localmente (o conjunto de dados é dividido em intervalos e a função é ajustada em cada um dos intervalos). Na representação gráfica o resultado é uma curva suave que funciona como modelo para os dados. Este processo destina-se a eliminar o enviesamento resultante das diferenças das intensidades da fluorescência inerentes aos dois fluoróforos utilizados, Cy3 e Cy5. Os valores de *RI* são ajustados, a partir deste processo, através de uma função de regressão de *RI* em *intensidade do controlo*×*intensidade do teste*. Neste contexto é designada por normalização *loess*.

	farray.Primary.Target	X1999V0053S	X1999V0906S	X1999V0980S	X1999V0993S	.	X1999V1076S
1	SP0001	1,263127	1,442144	1,054063	0,5641181	.	1,9067901
2	SP0002	0,809152	1,009282	1,015005	0,7695601	.	1,0737266
3	SP0005	1,039365	1,044655	1,299092	0,576497	.	1,7592398
4	SP0006	1,313541	1,294999	1,086401	0,8389911	.	1,3760739
5	SP0009	1,182621	1,326196	1,36174	0,8463099	.	1,8405826
6	SP0011	1,81294	1,858282	1,432585	0,9956734	.	1,9067901
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
3620	Spr2043	2,46929	6,39231	2,529483	1,2266	.	4,235609

Tabela 2.1. Matriz das médias de *RI* por gene (linhas) e estirpe (colunas).

Os valores de *RI* variam, teoricamente, entre valores próximos de 0 quando a intensidade do controlo é muito reduzida em relação à intensidade do teste (nos casos em que o gene se encontra presente e a respetiva fluorescência é muito elevada), e ∞, quando a intensidade referente à amostra é nula (caso em que o gene está ausente e apenas a amostra controlo hibridou com o oligonucleótido). *Rl* apresenta valores próximos de 1 quando as amostras de controlo e teste hibridaram em igual proporção com os oligonucleótidos (devido a iguais concentrações do teste e do controlo ou afinidades semelhantes com o oligonucleótido), superiores a 1 quando a amostra controlo está presente em maior proporção em relação à amostra teste, ou inferiores a 1 nos casos em que é a amostra teste que está presente em maior proporção na hibridação, ou tem maior afinidade. Uma vez que a amostra controlo é constituída por uma mistura de DNA das estirpes TIGR, R6 e G54, a intensidade de fluorescência do controlo é muito variável, visto que cada uma das três estirpes não contém todos os 3620 genes em estudo. No entanto, cada gene está presente em pelo menos uma delas.

A matriz de probabilidades de ausência foi obtida a partir destes dados através da aplicação do modelo de mistura Normal-Uniforme, metodologia desenvolvida por Dean e Raftery (2005). Este método, denominado *Normal Uniform Differential Gene Expression* (NUDGE), foi proposto com o fim de identificar genes diferencialmente expressos em estudos de expressão génica, e demonstrou elevada precisão (87%) ao ser aplicado na identificação de genes ausentes no estudo de Cardoso (2009). O algoritmo foi aplicado através da biblioteca NUDGE do R, com a diferença de que foram utilizadas as razões *Rl* em vez de $log_2 R$ (ver seção 1.1.3), e os valores de *Rl* não foram normalizados.

O modelo de mistura Normal-Uniforme permitiu modelar cada população de genes de cada

estirpe, em duas subpopulações, uma das quais pode ser ajustada a uma distribuição Normal – **genes presentes** – e a outra a uma distribuição Uniforme – **genes ausentes**. Este modelo de mistura permite calcular as probabilidades *a posteriori* de cada gene se encontrar ausente, que não necessitam de ser ajustadas para testes múltiplos. Cada uma das subpopulações é modelada pela sua densidade de probabilidade, logo, todo o conjunto de genes é modelado como uma mistura ponderada das duas densidades, em que os pesos correspondem às probabilidades *a priori* de pertença a cada um dos grupos.

O modelo é descrito pela seguinte equação:

$$x_i \overset{iid}{\sim} \pi N(x_i|\mu,\sigma^2)+(1-\pi)U_{[a,b]}(x_i), i=1,\ldots,N,$$

em que x_i é o valor de *RI* para o gene i, π é a probabilidade a *priori* de o gene estar presente, $N(x_i|\mu,\sigma^2)$ é uma distribuição Normal de valor médio μ e variância σ^2, $U_{[a,b]}(x_i)$ corresponde a uma distribuição Uniforme no intervalo $[a,b]$, e N é o número de genes.

Os parâmetros do modelo são estimados através do algoritmo EM. Uma vez que se desconhece que genes se encontram ausentes ou presentes, estes dados omissos são expressos pela variável Bernoulli Z, que toma valores $z_i=0$ se o gene i se encontrar ausente e $z_i=1$ se estiver presente ($i=1,\ldots,N$). O algoritmo é constituído por dois passos:

- *Estimação* (passo E) – os valores de z_i são calculados a partir das estimativas dos parâmetros π, μ e σ^2:

$$\hat{z}_i^{(k)}=\frac{(1-\hat{\pi}^{(k-1)})U_{[\hat{a},\hat{b}]}(x_i)}{\hat{\pi}^{(k-1)}N\left(x_i|\mu^{(\hat{k}-1)},\left(\sigma^{(\hat{k}-1)}\right)^2\right)+(1-\hat{\pi}^{(k-1)})U_{[\hat{a},\hat{b}]}(x_i)}, i=1,\ldots,N.$$

- *Maximização* (passo M) – as estimativas dos parâmetros π, μ e σ^2 são calculadas a partir dos valores z_i obtidos no passo E através dos seus estimadores de máxima verosimilhança:

$$\hat{\pi}^{(k)}=\frac{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)}{N},$$

$$\hat{\mu}^{(k)}=\frac{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)\times x_i}{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)},$$

$$\left(\hat{\sigma}^{(k)}\right)^2=\frac{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)\times\left(x_i-\hat{\mu}^{(k)}\right)^2}{\sum_{i=1}^{N}\left(1-\hat{z}_i^{(k)}\right)}.$$

A verosimilhança do modelo dadas as estimativas dos parâmetros na iteração k é dada por:

$$L\left(x;\hat{\pi}^{(k)},\hat{\mu}^{(k)},\left(\hat{\sigma}^{(k)}\right)^2\right)=\prod_{i=1}^{N}\left\{\hat{\pi}^{(k)}N\left(x_i;\hat{\mu}^{(k)},\left(\hat{\sigma}^{(k)}\right)^2\right)+\left(1-\hat{\pi}^{(k)}\right)U_{[\hat{a},\hat{b}]}\left(x_i\right)\right\} \quad .$$

(Dean e Raftery, 2005).

Os dois passos são iterados até ocorrer convergência. O critério de convergência é determinado a partir de um valor δ tal que: se na iteração k (k=1,...,j) sendo j o número máximo de iterações fixado à partida) $logL(k\text{-}1))\text{-}logL(k)<\delta$, as estimativas dos parâmetros e dos valores z_i serão respetivamente $\hat{\pi}^{(k)}$, $\hat{\mu}^{(k)}$, $\hat{\sigma}^{(k)}$ e $\hat{z}_i^{(k)}$ .O valor de δ é previamente fixado para que a convergência ocorra quando a diferença entre os valores estimados nas duas últimas iterações for suficientemente pequena para se considerar que estabilizaram.

As estimativas finais dos valores omissos z_i constituem as probabilidades *a posteriori* de cada gene se encontrar ausente, dadas as estimativas dos parâmetros.

As probabilidades de ausência foram assim obtidas para cada uma das estirpes, e deram origem à matriz de probabilidades de ausência que constituiu a base do presente estudo. Há que salientar que, sendo a precisão do método da ordem dos 87%, estima-se que cerca de 13% dos genes de cada estirpe se encontram erradamente classificados, o que em 3620 representa 471 genes.

A tabela seguinte apresenta uma parte da matriz de probabilidades de ausência obtida a aprtir da aplicação do modelo Normal-Uniforme:

	farray.Primary.Target	X1999V0053S	X1999V0906S	X1999V0980S	X1999V0993S	.	X1999V1076S
1	SP0001	0,00652167	0,007844032	0,01449502	0	.	0,009955927
2	SP0002	0	0,009065046	0,01522022	0	.	0,018770016
3	SP0005	0,007527396	0,008766592	0,01484777	0	.	0,010280396
4	SP0006	0,006456663	0,007722367	0,01407391	0	.	0,013467279
5	SP0009	0,006740518	0,007704896	0,016369	0	.	0,009922367
6	SP0011	0,009103562	0,011760284	0,01910977	0	.	0,00969504
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
3620	Spr2043	0,0473155	1	0,99073	0,238253	.	0,1022262

Tabela 2.2. Probabilidades de ausência de cada gene (linhas) em cada uma das estirpes (colunas).

Os dados biológicos e epidemiológicos referentes às estirpes em estudo encontram-se reunidos numa base de dados onde consta, para cada uma delas, o serotipo da cápsula, o clone (nos casos em que foi determinado), um número de identificação da estirpe, a razão das chances (*OR*) de causar doença invasiva respeitante ao serótipo e a classificação atribuída com base no *OR* (*invasive*, *colonization* e *neutral*). A tabela 2.3 representa uma parte destes dados. A distribuição das estirpes segundo a classe a que pertencem encontra-se representada na figura 2.1.

strain	sero	clone	st	OR sero	invasive	colonization	neutral
1999V0053S	33F		717	2,03457447	0	0	1
1999V0906S	3		260	2,63417659	1	0	0
1999V0980S	1		306	78,4037123	1	0	0
1999V0993S	9V		644	1,87986306	0	0	1
1999V1040S	14		557	2,03163322	1	0	0
1999V1216S	9N		66	9,82515991	1	0	0
2000V0189S	23F		81	0,4389325	0	1	0
2000V0324S	19A		276	0,89382716	0	0	1
2000V0527S	12B		1365	inf	1	0	0
2000V0626S	3		458	2,63417659	1	0	0
2000V0637S	12B		218	inf	1	0	0
2000V0731S	23F		338	0,4389325	0	1	0
2000V0734S	16F		414	0,04295593	0	1	0
2000V0926S	6B		1224	0,3032538	0	1	0
2000V1024S	14		15	2,03163322	1	0	0
2000V1277S	15B		1706	0,09430457	0	1	0
2001V0050S	6B	unique	273	0,3032538	0	1	0
2001V0240S	8	8-2	404	46,2857143	1	0	0
2001V0381S	19F	19F-1	177	0,10872162	0	1	0

Tabela 2.3. Excerto da tabela de dados referentes a cada uma das estirpes. A quinta coluna apresenta os valores de OR correspondentes ao serotipo de cada uma das estirpes. Os OR respeitantes às estirpes invasivas estão representados em fundo cinzento escuro, os que correspondem às estirpes neutras em cinzento claro e os que correspondem às colonizadoras em fundo branco. Inf - infinito.

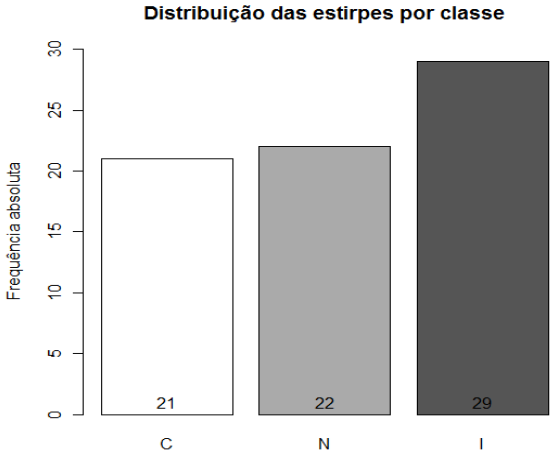


Figura 2.1. Gráfico de barras ilustrativo da frequência absoluta de cada uma das categorias atribuídas às estirpes em estudo. C – estirpes colonizadoras; N- estirpes neutras; N – estirpes invasivas.

As razões das chances foram calculadas a partir das suas proporções observadas em portadores saudáveis e com doença invasiva, com base no exemplo seguinte (o serotipo A representa um serotipo hipotético):

	Portadores Saudáveis	Doença Invasiva
Serótipo A	*a*	*b*
Serótipo Não A	*c*	*d*

$$OR_A = \frac{ad}{cb}$$

A classificação das estirpes em **invasivas**, **colonizadoras** ou **neutras** foi feita com base nos *OR* assim obtidos para cada um dos serótipos, com intervalos de confiança de 95%. As estirpes invasivas correspondem, assim, àquelas cujo serotipo apresenta um *OR* elevado (significativamente superior a 1), colonizadoras as que apresentam um valor de *OR* significativamente inferior a 1, e neutras aquelas para as quais o *OR* não foi significativamente diferente de 1 (Sá-Leão et al., 2011).

2.2.1. Análise Exploratória

Dado o elevado número de variáveis na amostra, tornou-se necessário organizar os dados da matriz de probabilidades de ausência de modo que fosse possível visualizar possíveis associações entre as estirpes e a presença/ausência dos genes. Para tal, a matriz de probabilidades foi transformada numa matriz indicadora de presença (também designada por matriz binária). Com base na aplicação do modelo Normal-Uniforme de Dean e Raftery (2005), que considerou diferencialmente expressos os genes que apresentavam probabilidades *a posteriori* superiores a 0.5, no presente estudo foram considerados ausentes os genes com probabilidades *a posteriori* de ausência iguais ou superiores a 0.5, e os restantes como genes presentes.

A classificação seguiu, assim, a seguinte regra:

- Genes ausentes: P(Ausente) $\geq 0{,}5 \rightarrow 0$
- Genes presentes: P(Ausente)$< 0{,}5 \rightarrow 1$

A tabela 2.4 representa um excerto da matriz resultante desta transformação.

	farray.Primary.Target	*X1999V0053S*	*X1999V0906S*	*X1999V0980S*	*X1999V0993S*	.	*X1999V1076S*
1	SP0001	1	1	1	1	.	1
2	SP0002	1	1	1	1	.	1
3	SP0005	1	1	1	1	.	1
4	SP0006	1	1	1	1	.	1
5	SP0009	1	1	1	1	.	1
6	SP0011	1	1	1	1	.	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
3620	Spr2043	1	0	0	1	.	1

Tabela 2.4. Matriz de probabilidades transformada em matriz indicadora de presença, onde 0 significa que o gene está ausente e 1 significa que está presente.

A figura 2.2 representa todo o conjunto de dados referentes aos genes, para cada uma das estirpes, com base na matriz de classificação binária.

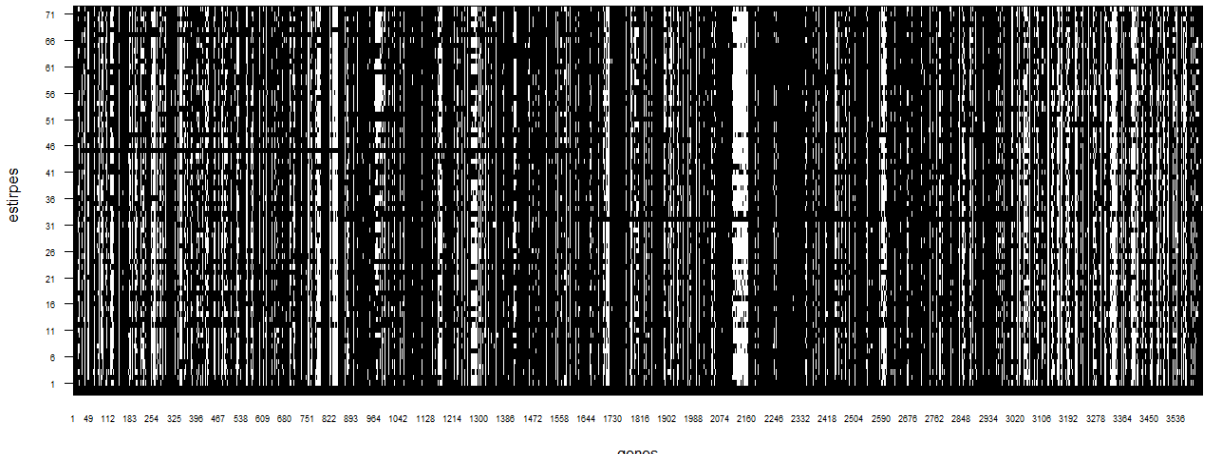


Figura 2.2. Genes presentes (preto) e ausentes (branco) em cada uma das 72 estirpes de Streptococcus pneumoniae.

A imagem ilustra a grande diversidade genética presente entre as estirpes em estudo. Na sua observação torna-se evidente que existe um elevado número de genes que estão presentes em todas as estirpes (zonas a preto), correspondendo muito provavelmente a parte do genoma essencial. São visíveis também algumas faixas brancas no gráfico, correspondentes a genes que estão ausentes na maior parte das estirpes, e zonas em mosaico onde não existe um padrão evidente.

O número de genes presentes em cada uma das estirpes, segundo o mesmo critério de classificação, encontra-se parcialmente representado na figura 2.3, e o resumo dos dados na tabela 2.5.

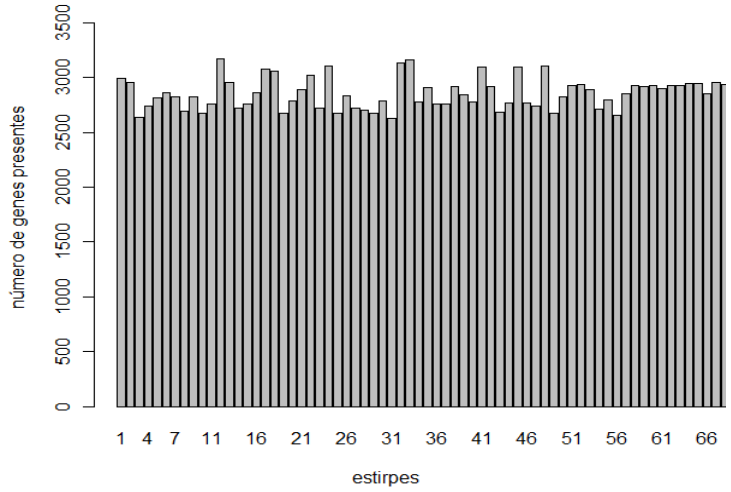


Figura 2.3. Gráfico de barras representativo do número total de genes presentes em algumas estirpes.

Média	2942
Mediana	2864
Desv. Padrão	139,1
Mínimo	2628
Máximo	3171

Tabela 2.5. Sumário dos dados referentes ao número de genes presentes em cada uma das estirpes.

Da observação destes resultados conclui-se que o número de genes pertencentes ao genoma essencial (presentes em todas as estirpes) não excederá o número mínimo de genes presentes em uma estirpe, que é de 2628. No entanto a estrutura de "mosaico" dos dados, observada no gráfico da figura 2.2, leva a crer que o número de genes pertencentes ao genoma essencial poderá ser bastante mais pequeno.

2.2.2 Genoma essencial e genoma acessório

Os genes potencialmente associados a doença invasiva não estarão presentes em todas as estirpes, tal como foi referido no capítulo anterior, pelo que a análise foi iniciada pela pesquisa dos genes potencialmente pertencentes ao genoma acessório, que serão os únicos que poderão dar a informação pretendida neste estudo.

Para definir os genes potencialmente pertencentes ao genoma essencial, por forma a

excluí-los da análise subsequente, foi utilizada a classificação binária baseada na matriz de probabilidades de ausência descrita anteriormente, de acordo com o seguinte procedimento:

1. Transformação da matriz de probabilidades de ausência *P.aus* numa matriz de probabilidades de presença *P.pre*.

$$1 - P.aus = P.pre$$

2. Transformação da matriz de probabilidades de presença na matriz indicadora de presença (referida em 2.2.1). Obteve-se assim uma matriz simplificada em que foi considerado ausente um gene i com probabilidade de pertença à estirpe j inferior a 0.5, e presente no caso de a sua probabilidade de pertença à respectiva estirpe ser igual ou superior a 0.5.

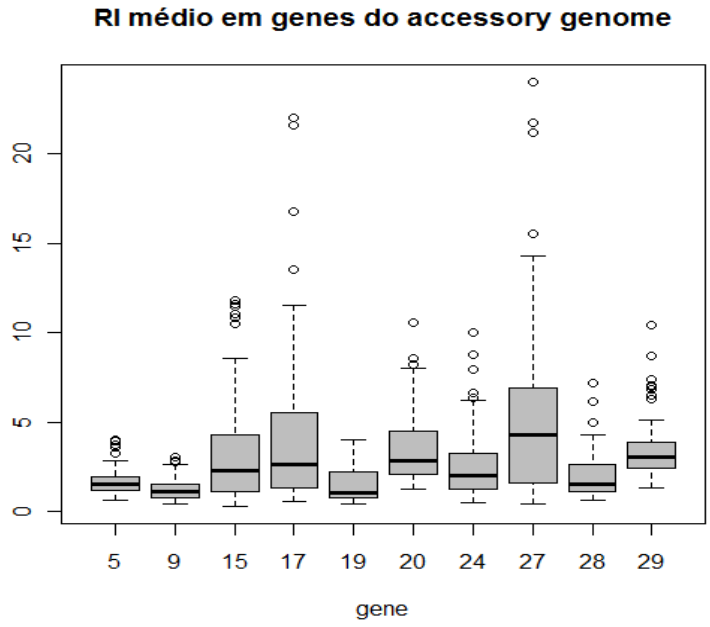
$$p_{ij} \geq 0{,}5 \Rightarrow g_{ij} = 1$$
$$p_{ij} < 0{,}5 \Rightarrow g_{ij} = 0$$

em que p_{ij} é a probabilidade de pertença do gene i à estirpe j, e g_{ij} a variável binária que indica se o gene está ausente (0) ou presente (1),

3. Exclusão dos genes potencialmente presentes em todas as estirpes, ou seja, todos aqueles em que $g_{ij}=1$ qualquer que seja a estirpe j.

Este procedimento permitiu reduzir a matriz de dados inicial, de dimensão 72×3620, a uma matriz de dimensão 72×1775, constituída pelos dados dos 1775 genes que se encontram ausentes em pelo menos uma estirpe, constituindo assim o genoma acessório dentro do universo dos 3620 genes do estudo.

Tomando as médias de *RI* obtidas para todos os genes em cada uma das estirpes, observa-se que, no caso dos genes pertencentes ao *core genome*, os valores obtidos encontram-se tendencialmente próximos de 1 e apresentam uma dispersão mais homogénea que no caso dos genes do genoma acessório, em que apresentam uma amplitude de variação maior e a sua distribuição parece ser mais heterogénea, o que não é de estranhar visto que se tratam de genes cujo padrão de presença nas várias estirpes é mais variável (figura 2.4).



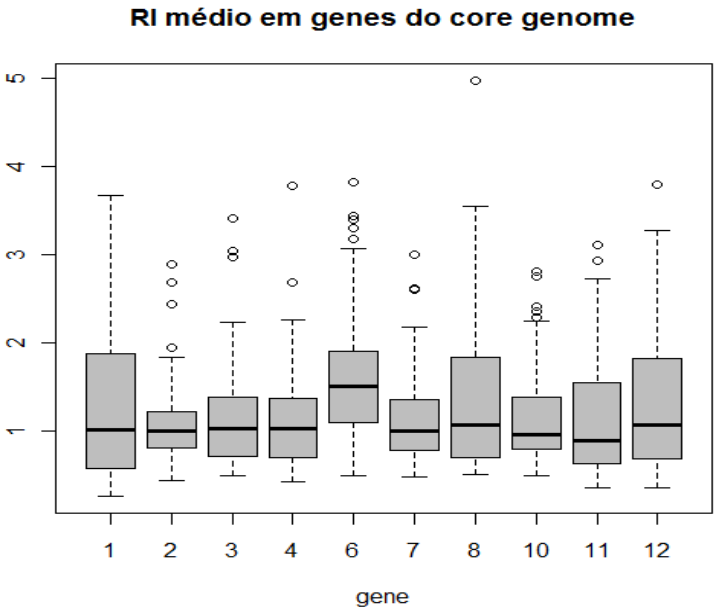


Figura 2.4. Caixas-com-bigodes representativas das RI médias de dez genes em todas as estirpes, para 10 genes do genoma acessório (accessory genome) *e 10 genes do genoma essencial* (core genome)*, respetivamente.*

Para que fosse possível visualizar possíveis padrões na nova matriz de dados foi construído um novo gráfico representativo da presença dos genes em cada uma das estirpes, desta vez apenas com os dados do genoma acessório (figura 2.5).

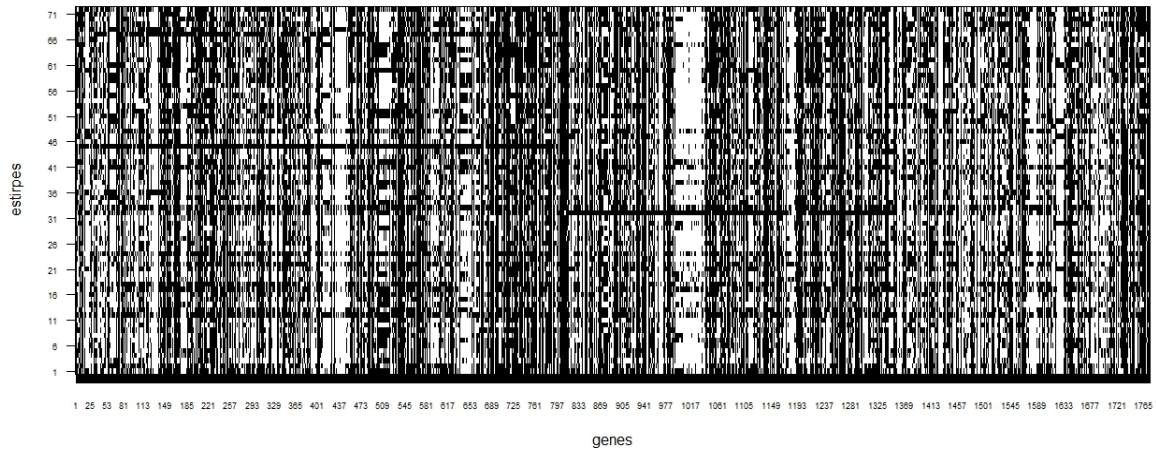


Figura 2.5. Genes potencialmente pertencentes ao genoma acessório em cada uma das 72 estirpes de Streptococcus pneumoniae. As zonas a preto representam os genes presentes.

A imagem mostra que o padrão de presença dos genes nas diferentes estirpes é bastante irregular, à exceção dos grupos de genes com comportamento semelhante já observados na figura 2.2.

Para mais facilmente visualizar possíveis relações entre a composição genética e a natureza das estirpes, estas foram agrupadas segundo as respetivas classificações quanto à capacidade invasiva (figura 2.6). Verifica-se que a distinção entre genes ausentes e presentes é muito mais clara no grupo das estirpes invasivas que nos grupos das colonizadoras e neutras, que visualmente apresentam um padrão mais semelhante entre si. Verifica-se também uma diferença mais marcada na faixa branca de genes ausentes já observada anteriormente, em que as proporções de genes presentes nesse grupo parecem ser maiores nas estirpes colonizadoras e neutras e quase nulas nas estirpes invasivas.

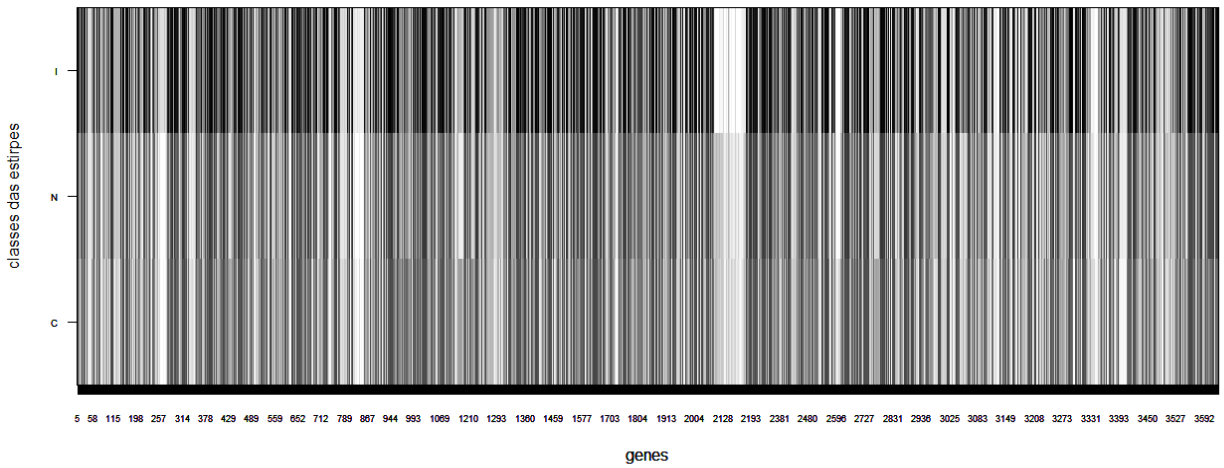


Figura 2.6. Representação gráfica das proporções dos genes potencialmente pertencentes ao genoma acessório em cada um dos grupos de estirpes (C-colonizadoras, N-Neutras, I-Invasivas). A proporção de presença em cada grupo é expressa por um gradiente de cor, em que o branco corresponde a 0 e o preto a 1 (ausente e presente em todas as estirpes, respetivamente).

2.3. Metodologias de Prospeção de Dados e a sua aplicação a dados de Hibridação Genómica Comparativa

Para cumprir o objetivo deste estudo, foram utilizadas metodologias de Prospeção de Dados (*Data Mining*) que permitissem avaliar a existência de associações entre a composição genética das diferentes estirpes e a classe a que pertencem (classificação *a priori*).

De acordo com Hand et al. (2001) em *Principles of Data Mining*, a Prospeção de Dados consiste na análise de conjuntos de dados observacionais, geralmente de grande dimensão, com a finalidade de encontrar relações anteriormente desconhecidas e resumir os dados de formas que se tornem simultaneamente compreensíveis e úteis ao seu utilizador.

O resumo dos dados e as relações encontradas que resultam do exercício da prospeção de dados são frequentemente designados por **modelos** ou **padrões**. Alguns exemplos incluem equações lineares, regras, agrupamentos, grafos, modelos em árvore, e padrões recorrentes em séries temporais.

A definição de prospeção de dados acima apresentada refere-se a dados observacionais, ao invés de dados experimentais, uma vez que estas metodologias não requerem os métodos de

amostragem específicos que são necessários à estatística (Hand et al., 2001). Neste estudo dispomos de dados experimentais, mas a sua dimensão e complexidade justificam a utilização destes métodos, que permitem encontrar possíveis padrões nos dados e selecionar, de entre as variáveis, aquelas que são mais informativas, o que neste caso significa encontrar os genes que melhor se correlacionam com o potencial invasivo de *Streptococcus pneumoniae*.

Os modelos testados têm como finalidade classificar as estirpes, da forma mais precisa possível, com base na sua composição genética e, de entre esses modelos, selecionar aquele que melhor permite classificar uma estirpe desconhecida.

Os processos de classificação pressupoem a existência de um conjunto de objetos a classificar, e um conjunto de variáveis ou atributos, referentes a cada um deles, das quais é extraída a informação que os vai permitir classificar. A classificação pode ser **não supervisionada**, quando é feita apenas a partir da informação fornecida pelas variáveis sem incluir conhecimento prévio acerca dos objectos, e **supervisionada**, nos casos em que, para além da informação extraída das variáveis, é utilizado conhecimento prévio acerca da natureza dos objetos. Neste caso a classificação é feita a partir da informação de um conjunto de objetos para os quais a classificação é já conhecida, denominado conjunto de treino. Neste estudo, em que as estirpes são os objetos e os genes as variáveis, foram testados várias metodologias de classificação: modelos descritivos na classificação supervisionada e modelos preditivos na classificação supervisionada, conforme se encontram abaixo descritos.

2.3.1. Métodos de classificação não supervisionada - Agrupamento não-hierárquico e agrupamento hierárquico

As metodologias de agrupamento hierárquico e não hierárquico consistem na decomposição de um conjunto de objetos em vários grupos, com base em medidas obtidas a partir das variáveis que os caracterizam, que exprimem o grau de proximidade entre os mesmos. Obtêm-se assim grupos em que os objetos neles incluídos são mais semelhantes entre si que em relação aos que pertencem a outros grupos. Existem vários métodos de particionamento e várias medidas de dissemelhança que podem ser utilizados, e a sua escolha deve ser adequada à natureza dos dados que estão a ser analisados.

2.3.1.1 Medidas de dissemelhança

Para que seja possível comparar objetos entre si e construir uma classificação em diferentes grupos, é necessário dispôr de medidas de semelhança ou dissemelhança entre os objetos. Essas medidas são obtidas a partir dos valores que cada variável assume para cada objeto, que constituem vetores de dimensão igual ao número de variáveis disponíveis. Normalmente nos cálculos são utilizadas as dissemelhanças, designando-se por $d(i,j)$ a dissemelhança entre os objetos i e j. A semelhança entre eles será designada por $s(i,j)$, obtida por conversão: $s(i,j)= 1-d(i,j)$.

Os termos **distância** e **métrica** são frequentemente usados neste contexto. O termo distância é utilizado informalmente para designar uma medida de dissemelhança resultante das características que descrevem os objetos, como a distância euclideana. Uma métrica é uma medida de dissemelhança que satisfaz os seguintes critérios:

1. $d(i,j) \geq 0 \ \forall \ i,j$ e $d(i,j)=0$ sse $i=j$
2. $d(i,j) = d(j,i) \ \forall \ i,j$
3. $d(i,j) \leq d(i,k) + d(k,j) \ \forall \ i,j,k$ (Desigualdade triangular).

Existem diversas medidas de dissemelhança que podem ser utilizadas, dependendo da natureza das variáveis (que podem ser categóricas, ordinais ou escalares), e do peso que se pretende atribuir a cada variável.

No caso em que as variáveis são escalares a medida mais simples é a distância euclideana, e existem outras medidas conhecidas que constituem variações da mesma. A escolha da medida a adotar deve ser feita de forma a adequar-se à natureza dos dados.

Supondo que se dispõe de n objetos para cada um dos quais foram obtidas p medidas. O vetor de observações do i-ésimo objeto será $x(i) =(x_1(i), x_2(i),..., x_p(i))$, $1 \leq i \leq p$, em que o valor da k-ésima variável para o i-ésimo objeto é $x_k(i)$.

Medidas de distância aplicáveis a variáveis quantitativas

- *Distância euclideana*

A distância euclideana entre o objeto i e o objeto j é obtida a partir de:

$$d_E(i,j)=\sqrt{\sum_{k=i}^{p}\left(x_k(i)-x_k(j)\right)^2} \quad .$$

Esta medida assume que as variáveis são comensuráveis, ou seja, têm escalas e limites de variação semelhantes. Caso isso não se verifique as variáveis devem ser standartizadas. Além disso, todas as variáveis têm o mesmo peso na medida obtida, pois não é incluído nenhum coeficiente de ponderação, e as variáveis são tratadas como sendo independentes.

No caso de se pretender atribuir diferentes pesos às variáveis, existem as seguintes alternativas:

- *Distância euclideana ponderada*

$$d_{WE}(i,j)=\sqrt{\sum_{k=i}^{p} w_k\left(x_k(i)-x_k(j)\right)^2} \quad ,$$

em que w_k é a k-ésima corresponde a um vetor w de pesos a atribuir às p variáveis, conforme a sua importância relativa.

- *Distância de Mahalanobis*

$$d_{MH}(i,j)=\sqrt{\left(x_k(i)-x_k(j)\right)^T \Sigma^{-1}\left(x_k(i)-x_k(j)\right)} \quad ,$$

sendo Σ^{-1} a inversa da matriz de covariância $p \times p$.

Esta medida é adequada quando existem variáveis fortemente correlacionadas, sendo por isso aconselhada sempre que $\Sigma^{-1} \neq I$, de forma a introduzir um coeficiente de ponderação que atribua menos peso às variáveis que se encontram fortemente correlacionadas e maior peso às variáveis não correlacionadas, o que permite compensar o efeito de redundância resultante da introdução de variáveis que fornecem o mesmo tipo de informação Quando a matriz de covariâncias não é invertível não é possível calcular esta distância.

- *Distância de Minkowski ou métrica L_λ*

$$d_{Mk}(i,j)=\left(\sum_{k=i}^{p}\left(x_k(i)-x_k(j)\right)^{\lambda}\right)^{\frac{1}{\lambda}}$$

Trata-se de uma generalização da distância euclideana. No caso em que $\lambda=1$, designa-se por **distância de Manhattan**, e a distância euclideana corresponde ao caso em que $\lambda=2$.

- *Distância com base em correlação*

Medida de dissemelhança obtida a partir do cálculo da correlação, designada doravante por **distância correlação.** Trata-se de uma medida considerada neste estudo com o propósito de efetuar a classificação hierárquica dos genes a partir das suas probabilidades de pertença, considerando mais próximos genes com valores de correlação próximos de 1 (correlação máxima) e mais dissemelhantes aqueles cujos valores de correlação se aproximam de -1.
Esta medida é calculada através do seguinte processo:

1. Cálculo da matriz de correlação através de coeficiente de correlação de Pearson para cada par de variáveis x e y, *em que* $i=1,...,n$ correspondem aos objetos.

$$r_{x,y}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

2. Conversão da matriz de correlações numa matriz de dissemelhanças. A correlação entre x *e* y varia entre -1, se se encontram totalmente correlacionados mas a proporcionalidade é inversa, e 1, quando a correlação é máxima e são diretamente proporcionais. Ao interpretar estes valores como uma medida proporcional à dissemelhança entre os genes, há que ter em conta que os valores mínimos de correlação, próximos de -1, refletem uma maior dissemelhança enquanto que os valores máximos refletem menor dissemelhança. A transformação é, assim, efetuada de acordo com a fórmula seguinte:

$$d_{Corr}(x,y)=0{,}5(1-r_{x,y})$$

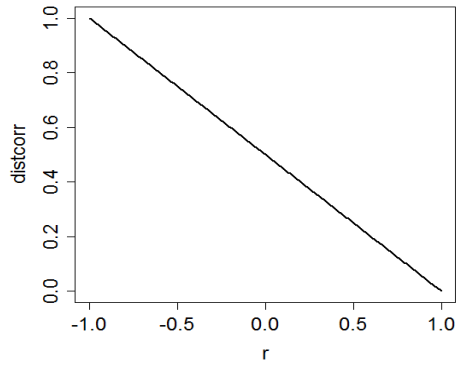


Figura 2.7. Gráfico representativo da conversão dos valores de correlação em dissemelhanças. r=-1 corresponde à distância máxima (1), enquanto que r=1 corresponde à distância mínima (0).

Medidas de semelhança aplicáveis a variáveis categóricas

Nos casos em que os dados são categóricos, existem medidas mais adequadas, que se baseiam na contagem do número de variáveis em que dois objetos i e j tomam o mesmo valor ou valores diferentes O número de variáveis para as quais ambos, por exemplo, tomam o valor 0 é designado por $n_{0,0}(i,j)$. Por uma questão de simplificação de escrita é omitida a referência ao par (i,j) sempre que for evidente a identificação dos objetos.

No caso particular em que os dados são binários, para cada par de objetos obtêm-se os seguintes valores:

	j=0	j=1
i=0	$n_{0,0}$	$n_{0,1}$
i=1	$n_{1,0}$	$n_{1,1}$

Tabela 2.6. Valores que um par de objetos i e j podem tomar de acordo com um conjunto de variáveis binárias.

$n_{0,0}$ é o número de variáveis para as quais os objetos i e j tomam ambos o valor de 0, $n_{0,1}$ o número de variáveis para as quais i=0 e j=1, $n_{1,0}$ o número de variáveis em que i=1 e j=0, e $n_{1,1}$ o número de variáveis para as quais i e j tomam ambos o valor de 1.

- *Coeficiente de concordância simples*

É a medida de semelhança mais simples para dados binários, baseando-se no número de observações concordantes em cada par de objetos.

$$CCS(i,j)=\frac{n_{0,0}+n_{1,1}}{n_{0,0}+n_{0,1}+n_{1,0}+n_{1,1}}$$

O coeficiente de concordância simples (CCS) pode ser transformado numa medida de dissemelhança através da operação:

$$d_{CCS}(i,j)=1-CCS(i,j)$$

Quando o número de concordâncias numa das categorias não constitui uma informação relevante no estudo ou se se pretende atribuir maior peso a uma das categorias, em alternativa, podem ser utilizados os coeficientes de *Jaccard* e *Dice*, respetivamente.

- *Coeficiente de Jaccard*

$$CJ(i,j)=\frac{n_{1,1}}{n_{0,1}+n_{1,0}+n_{1,1}}$$

- *Coeficiente de Dice*

$$CD(i,j)=\frac{2\,n_{1,1}}{n_{0,1}+n_{1,0}+n_{1,1}}$$

2.3.1.2. Algoritmos baseados em particionamento -agrupamento não-hierárquico

Nos algoritmos baseados em métodos de particionamento o objetivo é dividir o conjunto de dados em K subconjuntos de forma que os pontos incluídos em cada subconjunto sejam o mais homogéneos possível: partindo do conjunto de n pontos $D=\{x(1),..., x(n)\}$, a tarefa é encontrar K *clusters* $C=\{C_1,...C_K\}$ de forma que cada ponto $x(i)$ seja incluído num único *cluster* C_K. A homogeneidade dos *clusters* resultantes é avaliada por uma função *score*, que é otimizada quando a variação dentro de cada *cluster* é minimizada relativamente à variação entre os *clusters*.

A função *score* pode ser obtida a partir da determinação do **centróide** ou média dos pontos do *cluster*, sendo, no processo de classificação, minimizada ou maximizada (dependendo da função escolhida) através de algoritmos iterativos.

- ***Algoritmo K*-means**

Neste algoritmo, tal como em outros algoritmos baseados em particionamento, o número de *clusters K* é fixado à partida.

Sendo $d(x,y)$ a distância entre os pontos $x, y \in C_K$, os centróides de cada cluster, r_K, são obtidos calculando a média aritmética entre os seus pontos:

$$r_k = \frac{1}{n_k} \sum_{x \in C k} x,$$

em que n_K é o número de pontos pertencentes ao k-ésimo *cluster*.

A variação intra-cluster $wc(C)$ e a variação inter-cluster $bc(C)$ (*within cluster* e *between cluster*) são depois calculadas, e a partir das mesmas é possível avaliar a qualidade das partições:

$$wc(C) = \sum_{k=1}^{K} wc(C_k) = \sum_{k=1}^{K} \sum_{x(i) \in C_k} d(x, r_k)^2$$

$$bc(C) = \sum_{1 \leq j < k \leq K} d(r_j, r_k)^2$$

A função *score* de C é definida como uma combinação monótona dos fatores $wc(C)$ e $bc(C)$, como por exemplo o rácio $bc(C)/wc(C)$.

Os K centros iniciais são escolhidos aleatoriamente. De seguida, cada ponto é alocado ao *cluster* cujo centro esteja mais próximo, de acordo com a distância euclideana. Os centros dos *clusters* são então recalculados, e o processo é repetido até que nenhum ponto deva ser realocado.

2.3.1.3. Agrupamento hierárquico

Os algoritmos de agrupamento hierárquico iniciam-se a partir de todo o conjunto de dados, constituído pelos pontos que representam as distâncias entre cada objeto, e gradualmente dividem, ou aglomeram, os pontos, recalculando em cada passo as distâncias entre os grupos recém-formados, até todos os pontos se unirem na mesma estrutura ou cada ponto se encontrar individualizado. Esta estrutura é representada através de um diagrama em árvore, designado por **dendograma**, no qual as divisões são representadas, e o tamanho de cada ramo é proporcional às distâncias entre os pontos ou grupos que nele se encontram incluídos.

Os algoritmos podem ser **aglomerativos**, quando unem em cada passo, os *clusters* mais próximos, ou **divisivos**, no caso em que todo o conjunto de pontos constitui inicialmente um único *cluster* e em cada passo é efetuada uma divisão em novos *clusters*, progressivamente, até à fase em que cada *cluster* contém um só ponto.

Os métodos aglomerativos são os mais largamente utilizados e mais simples de implementar. Nestes métodos as distâncias entre os *clusters* são recalculadas após cada passo do algoritmo, e com base nas novas distâncias é adicionado aos grupos recém-formados o grupo, ou ponto, cuja distância é menor. Existem várias formas de recalcular estas distâncias, das quais se destacam o método do **vizinho mais próximo** ou **ligação simples** e do **vizinho mais afastado** ou **ligação completa**.

- ***Ligação simples ou método do vizinho mais próximo***

Considera-se a distância entre dois *clusters* C_i e C_j como a distância entre os dois pontos x e y mais próximos tais que x pertence a C_i e y pertence a C_j.

$$D_{sl}(C_i, C_j) = \min_{x,y} \left\{ d(x, y) \middle| x \in C_i, y \in C_j \right\}$$

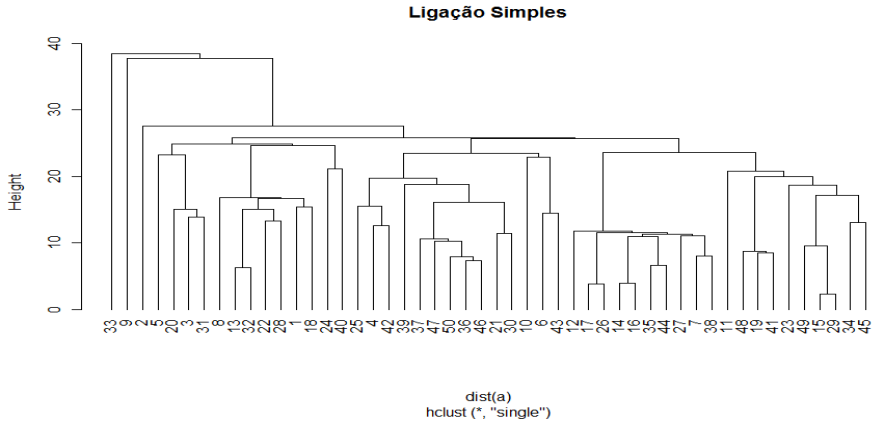


Figura 2.8 Exemplo de dendograma construído através do método de ligação simples.

- ***Ligação completa ou método do vizinho mais afastado***

Neste caso a distância entre os clusters *Ci* e *Cj* corresponde à distância entre os dois pontos *x* e *y* mais afastados tais que *x* pertence a *Ci* e *y* pertence a *Cj*.

$$D_{cl}(C_i, C_j) = \max_{x,y} \{ d(x,y) | x \in C_i, y \in C_j \}$$

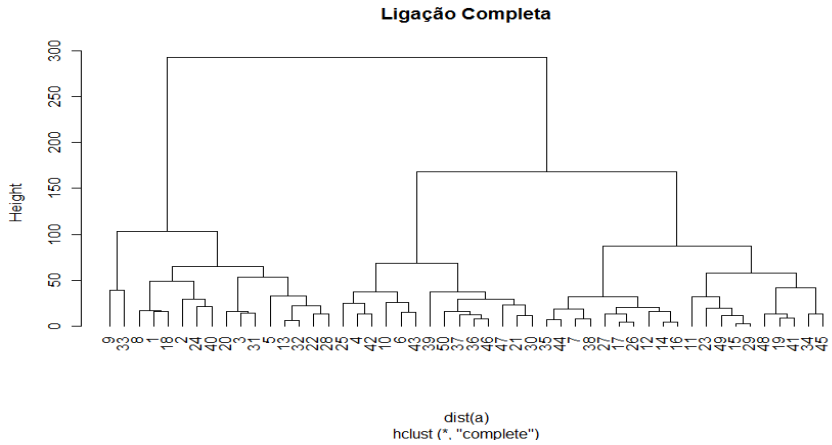


Figura 2.9. Exemplo de dendograma construído através do método de ligação completa.

A ligação simples, uma vez que tende a aglomerar objetos que se encontram a menores distâncias, tende a produzir *clusters* alongados, e é mais suscetível de produzir encadeamentos no dendograma tornando difícil a indentificação de grupos.

A ligação completa é um método mais adequado quando se pretende identificar grupos mais compactos nos dados, uma vez que ao aglomerar os pontos mais afastados ao longo do algoritmo, tende a formar *clusters* esféricos e bem individualizados. Devido a essa propriedade, o agrupamento hierárquico através de ligação completa foi o método de classificação escolhido neste estudo na tentativa de identificar grupos nas estirpes de *Streptococcus pneumoniae* com base na sua composição genética, e posteriormente para a seleção de genes a utilizar na construção do modelo em árvore.

Ao contrário do que sucede no agrupamento não-hierárquico, no agrupamento hierárquico não é necessário definir inicialmente um número de *clusters* que se pretende formar. Se o objetivo é encontrar agrupamentos entre os objetos, é escolhido um ponto de corte na altura do dendograma que, de acordo com o objetivo da classificação, pode resultar de uma escolha prévia do número de *clusters* que se pretende ou da visualização dos agrupamentos naturais que se formaram. Na figura 2.10 é apresentado o dendograma da figura 2.9 após a divisão dos objetos em três grupos.

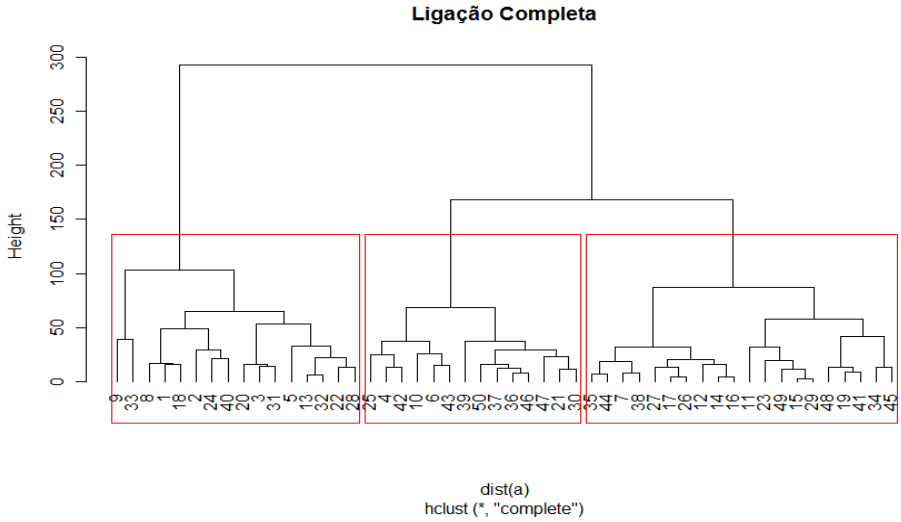


Figura 2.10. No dendograma foi escolhido um ponto de corte a uma altura próxima de 150 de forma a individualizar três grupos de objetos classificados.

2.3.2. Classificação supervisionada - Exploração de dados através de modelos em árvore

Os modelos preditivos de classificação podem ser vistos como a interpretação de associações a partir de um conjunto de variáveis de entrada (*input*) X relativas a um escalar y que corresponde à variável resposta. Esta variável resposta Y é, nos modelos de classificação uma variável categórica. A variável a predizer a partir das medidas obtidas, normalmente denominada **variável de classe** e designada por C toma valores no conjunto $\{c_1,...,c_m\}$. As variáveis medidas ou observadas $X_1,...,X_p$ são normalmente designadas por atributos, variáveis explicativas, ou variáveis de entrada. X é um vetor p-dimensional e as variáveis que o constituem podem ser de qualquer tipo (categóricas, ordinais ou quantitativas, por exemplo). $x_j(i)$ corresponde assim à j-ésima componente da i-ésima variável de entrada, em que $1 \leq i \leq n$ e $1 \leq j \leq p$ (Hand, 2001).

O princípio básico dos modelos em árvore é a partição, de forma recursiva, do espaço formado pelas p variáveis de entrada, de forma a maximizar o nível de "pureza" das classes, de forma que a maioria dos pontos de cada subconjunto resultante da partição pertença a uma única classe. Assim se, por exemplo, tomarmos três variáveis de entrada x, y e z, a variável x pode ser particionada de forma que o espaço constituído pelas variáveis de entrada fique dividido em duas células. Cada uma destas é particionada novamente de acordo com os valores de x, de y ou z. O processo é repetido até cada ramo terminal da árvore conter pontos pertencentes a uma só classe. Para predizer a classe a que pertence um novo objeto cujos valores das variáveis de entrada são conhecidos, a árvore é percorrida no sentido descendente, alocando o novo objeto ao ramo apropriado de acordo com os valores que toma em cada uma das variáveis que presidem a cada divisão na árvore.

Os modelos em árvore têm várias propriedades que os tornam uma ferramenta atrativa: são fáceis de compreender e de explicar, permitem incluir simultaneamente variáveis de naturezas distintas, e são flexíveis e eficientes como ferramenta de predição. No entanto, devido à forma como são construídas, pode acontecer levarem a partições subótimas no espaço das variáveis de entrada (Hand, 2001).

A validação do modelo permite, posteriormente, avaliar a sua qualidade preditiva de forma a que possa ser otimizado.

Ao efetuar a partição do conjunto de dados segundo os valores que uma variável X_j quantitativa toma, é necessário definir um valor de charneira (*threshold*) para cada variável, que constitua o critério da partição. Por exemplo, se a variável for binária o critério de partição obedece à condição $x_j(i)=0$ ou $x_j(i)=1$, mas se a variável for contínua (por exemplo $0 \leq x_j(i) \leq 1$), é necessário definir um valor t tal que o critério seja $x_j(i) \leq t$ ou $x_j(i) > t$.

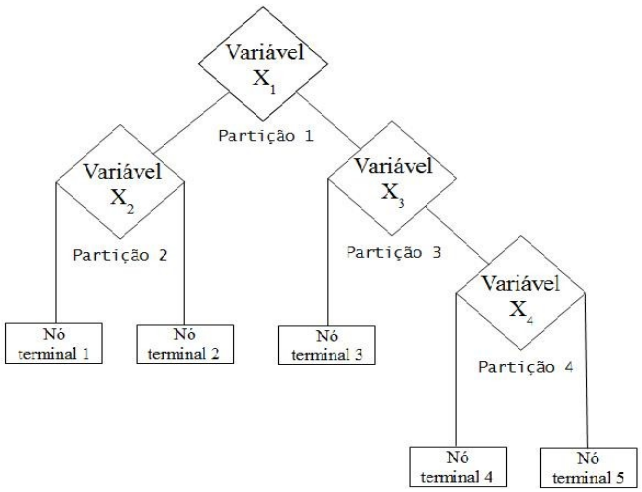


Figura 2.11. Esquema genérico de um modelo em árvore.

A escolha da variável que preside a cada uma das partições é feita a partir da informação fornecida por uma função *score* calculada para cada uma das variáveis, cujo valor reflete a associação existente entre as mesmas e a classificação conhecida dos objetos que fazem parte do conjunto cuja informação é utilizada na construção do classificador – **conjunto de treino**. Existem vários critérios para a determinação desta função *score*, dos quais se destacam a **taxa de erros da classificação**, o **índice de Gini** e a **entropia**. Segundo Hastie et al. (2009), a entropia e o índice de Gini são mais sensíveis às mudanças de probabilidades nos nós da árvore do que a taxa de erros da classificação, sendo por isso preferíveis, pois mais facilmente produzem partições que resultam em nós mais puros.

- ***Taxa de erro da classificação***

A taxa de erro da classificação estima a proporção de objetos incorretamente classificados ao efetuar uma partição segundo um determinado valor de charneira de uma variável X_j, num dado nó *t* da árvore. Esta taxa é calculada através de:

$$Erro\,da\,classificação(t) = 1 - \max_k p_k \quad ,$$

em que *k* corresponde a uma dada classe pertencentes ao conjunto de classes *C* (*k*=1,...*m*), e p_k à proporção de objetos que a ela são atribuídos no nó da árvore. A variável que dá origem à melhor partição será aquela para a qual o erro da classificação é menor.

- ***Índice de Gini***

O índice de Gini num dado nó *t* da árvore corresponde a:

$$Gini(t) = 1 - \sum_{C_k} p_k^{\,2} \quad ,$$

em que p_k são as proporções em que os objetos das diferentes classes C_k se encontram no nó da árvore para o qual é calculado este índice. Desigualdades maiores entre as proporções das diferentes classes dão origem a um índice de Gini mais baixo, verificando-se o contrário quando as proporções estão mais equilibradas. Desse modo a variável que dá origem à melhor partição será aquela que leva à maior redução do índice de Gini nos ramos que dela resultam.

- ***Entropia***

De acordo com Bramer (2007) em *Principles of Data Mining*, a experiência mostra que a entropia dá geralmente origem a árvores com menos ramos do que outros critérios de seleção dos atributos, e a resultados mais precisos, embora não haja garantias de infalibilidade. A entropia foi, assim, o critério escolhido neste estudo para a elaboração dos modelos de classificação.

Na construção de um modelo de classificação para uma variável desfecho (categórica) *C*, a entropia de um teste *T*, (onde *T* consiste no teste que verifica se $Xj > t$ onde *t* é o valor de charneira) define-se como a entropia média após a realização do teste:

$$H(C|T) = p(T=0)H(C|T=0) + p(T=1)H(C|T=1) \quad ,$$

em que a $H(C|T=l)$ é a entropia condicional a $T=l$.

$$H(C|T=l)=-\sum_{c_k} p(c_k|T=l)\log_2 p(c_k|T=l) \quad .$$

C_k, k= 1,...,m representam as classes da variável desfecho C, l=0,1 e, dado um valor de charneira t, T=1 *se* $X_j > t$, *e* T=0 *se* $Xj \leq t$. Nos casos em que $p(c_k|T=l)=0$ atribui-se o valor 0 à expressão $\log_2 p(c_k|T=l)$.

A entropia média define-se como a medida de incerteza associada a cada ramo pesada pelas respetivas probabilidades (Antunes, 2009; Hand, 2001).

A entropia do sistema $H(C|T)$ corresponde, assim, a uma medida de impureza associada a C, uma vez em que o seu valor é tanto maior quanto menores são as diferenças de proporcionalidade entre as classes C_k. Se existirem m desfechos possíveis (classes), o valor máximo que a entropia pode tomar corresponderá a $log_2 m$. Assim, no caso particular em que existem dois desfechos possíveis, e as suas proporções são respetivamente p e 1-p o valor máximo possível de $H(C|T)$ é $log_2 2$=1, que corresponde ao caso em que a proporcionalidade de objetos pertencentes a cada uma das classes é igual (figura 2.12).

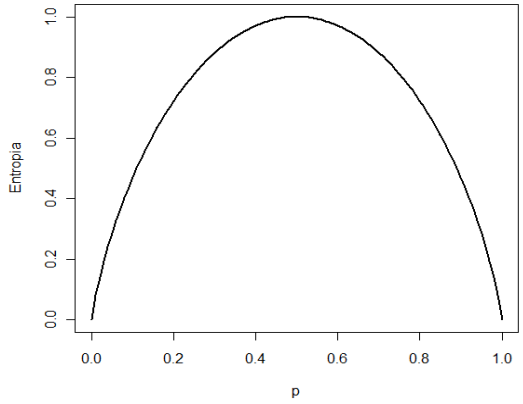


Figura 2.12 Valores de entropia num sistema com duas classes, como função de p. O valor máximo de entropia observa-se quando p=0,5.

No caso de X_j ser uma variável com k_j níveis, T toma valores em $\{1,...,k_j\}$ e, consequentemente, l=1,..., k_j. Assim, a entropia do sistema considerando a introdução da variável X_j é obtida a partir de:

$$H(C|T)=\sum_{l=1}^{k_j} p(T=l)H(C|T=l)$$

Ao dividir o conjunto de dados em subconjuntos nos quais a maioria dos objetos pertence a uma classe, a entropia em cada um desses subconjuntos (ramos) é reduzida. Assim, todo o processo é conduzido no sentido de reduzir a entropia média após cada particionamento.
A redução da entropia resultante da inclusão de um teste sobre a variável *Xj* designa-se por **ganho** e o seu valor é determinado por:

$$Ganho(H(C), X_j) = H(C) - H(C|X_j) \quad ,$$

em que $H(C)$ é a entropia do sistema e

$$H(C|X_j) = \sum_{l=1}^{k_j} p(X_j = l)\, H(C|X_j = l)$$

Assim, antes de ser efetuada a partição de um conjunto de dados, são calculadas a entropia do sistema e as entropias condicionais a $T=l$ para cada uma das variáveis X_j. A partir desses dados são obtidos os valores do ganho correspondentes à inclusão de um teste sobre cada uma das variáveis. A variável cujo valor de ganho é mais elevado é, assim, aquela cuja partição dá origem a dois subconjuntos de dados cuja entropia é mais reduzida, sendo escolhida como critério da partição do conjunto original. O procedimento pode ser repetido, até os nós terminais conterem apenas observações da mesma classe, como consequência do decréscimo da entropia em cada passo.

- ***Construção do algoritmo de classificação***

Ao construir um classificador até se esgotarem as possibilidades de partição poderão aparecer ramos com poucas observações ou mesmo uma só observação do conjunto de treino. Estas partições não resultam geralmente de uma associação da variável com as classes dos objetos, sendo muitas vezes fruto do acaso, uma vez que os objetos a classificar se encontram em número reduzido. Isto pode resultar num modelo sobreajustado, o que constitui uma desvantagem quando a finalidade é predizer o valor da classe de uma nova observação. Neste caso o modelo pode ser alterado, normalmente através da união dos ramos terminais com número reduzido de observações, a um nível que permita classificar com maior precisão um novo objeto. O classificador poderá assim adquirir melhor capacidade preditiva, visto que são eliminadas variáveis não informativas. Esta estratégia denomina-se ***pruning***, uma vez que elimina ramos terminais, por analogia com a poda de uma árvore.

O algoritmo CART (*Classification and Regression Trees*) é um procedimento estatístico largamente utilizado para produzir modelos de classificação e de regressão com estrutura em árvore (Hand, 2001). Este algoritmo foi desenvolvido por Breiman et al. e publicado em 1984. Com base no algoritmo CART foram desenvolvidos diversos programas informáticos de construção de modelos em árvore, nomeadamente a biblioteca RPART, utilizada no presente estudo.

No que se refere aos modelos de classificação, o algoritmo CART opera através da escolha da variável que origina uma melhor partição do conjunto de dados em dois grupos o mais homogéneos possível. Existem vários critérios que podem ser utilizados na escolha dessa variável, sendo um deles a redução da entropia (**ganho**) correspondente à sua inclusão, descrito anteriormente. Este algoritmo permite construit árvores de diversas dimensões consoante o nível de *pruning* escolhido (definido pelo número mínimo de objetos em cada ramo terminal).

Na construção da árvore é incluída uma única variável em cada passo, o que permite trabalhar com um grande número de variáveis. No entanto, o poder representativo da estrutura em árvore é um tanto limitado, uma vez que as regiões de decisão da classificação se encontram restringidas a estruturas em hiper-retângulo, com fronteiras paralelas aos eixos da variável de entrada (Hand, 2001), que nem sempre representam rigorosamente as fronteiras entre os objetos pertencentes a diferentes classes.

A função *score* utilizada para medir a capacidade preditiva de um modelo em árvore é uma função **perda** que se baseia-se no número de objetos incorretamente classificados pelo modelo:

$$\sum_{i=1}^{n} C\left(y(i), \hat{y}(i)\right) \quad ,$$

em que $C\left(y(i), \hat{y}(i)\right)$ é a perda obtida (positiva) quando a classificação predita para o i-ésimo vetor de variáveis, $y(i)$, corresponde a $\hat{y}(i)$, o que significa que o i-ésimo objeto foi corretamente classificado. Geralmente, C é representada por uma matriz $m \times m$, em que m corresponde ao número de classes. Assume-se a existência de uma perda igual a 1 sempre que $y(i) \neq \hat{y}(i)$, e igual a 0 no caso oposto (Hand, 2001).

A matriz $m \times m$ é denominada matriz de perda (*loss matrix*), e permite obter a **taxa de erros da classificação** (definida anteriormente) Esta taxa é utilizada para estimar a capacidade preditiva dos modelos.

- ***Avaliação da capacidade preditiva do modelo***

A função perda é estimada através de uma metodologia denominada **validação cruzada**, que consiste em particionar aleatoriamente o conjunto de dados em dois subconjuntos, o primeiro que funcionará como **conjunto de treino *L*** a partir do qual é construído o modelo, e o segundo, o **conjunto de teste *T***, cujos objetos são classificados através do modelo permitindo calcular a função *score* (perda). Este particionamento é repetido múltiplas vezes em diferentes subconjuntos de treino e de teste, e é calculada a média das taxas de erros da classificação, que permite avaliar como o modelo se comporta com novos dados cuja classificação é desconhecida.

Existem várias estratégias de efetuar o processo de validação cruzada, das quais se destacam as seguintes:

- **Validação cruzada em blocos** – o conjunto de dados é dividido em vários subconjuntos da mesma dimensão, ou de dimensões aproximadas (blocos). Sendo $K = 1,...,k$ o número de blocos e d a dimensão dos blocos fixada à partida (número de objetos em cada bloco), o conjunto de treino L constituído por k-1 blocos selecionados aleatoriamente constituirá o conjunto de dados sobre o qual será construído o modelo, e o conjunto de teste T será formado pelo bloco que resta. Os objetos incluídos em T são classificados segundo o modelo construído, e é estimada a taxa de erros da classificação. O processo é repetido até todos os k blocos serem classificados à custa do modelo construído pelos restantes dados, e por fim é estimada a taxa de erros do classificador, através da média das taxas obtidas na classificação de cada um dos blocos.

- **Validação cruzada *leave one out*** - O procedimento é o mesmo, com a diferença de que neste caso cada bloco é constituído por apenas um objeto ($k=n$ e $d=1$). Assim, num conjunto de dados com i objetos são construídos i modelos de classificação em que o conjunto de treino tem dimensão i-1 e o conjunto de teste é constituído apenas pelo vetor de variáveis correspondente ao objeto que resta. Neste caso particular a função *score* em cada bloco é uma variável binária que assume o valor 1 se $y(i) \neq \hat{y}(i)$ e 0 no caso oposto.

Neste estudo, foi escolhida a validação cruzada *leave one out*, por se considerar que um modelo construído com base num maior número de observações (neste caso 71 uma vez que i=72) constitui uma melhor aproximação à realidade do que um modelo construído com base num menor número observações, o que acontece sempre que K é reduzido.
Na construção dos modelos foram utilizados e testados vários algoritmos, que diferem na forma como as variáveis foram selecionadas e na forma como o particionamento foi efetuado.

2.3.3. Critério da seleção das variáveis e algoritmos de classificação

A seleção das variáveis para iniciar a construção dos modelos foi feita com base no cálculo da entropia do sistema e entropia condicional às 1775 variáveis do estudo. Foi assim obtido o ganho (redução da entropia) correspondente à inclusão de cada uma das variáveis no modelo, que serviu de ponto de partida para a seleção das variáveis mais informativas. Os cálculos foram efetuados a partir da matriz indicadora de presença e da classificação das estirpes, através de funções desenvolvidas no R para o efeito, tendo por base as fórmulas do cálculo da entropia do sistema, entropia condicional e ganho (apêndice A).

2.3.3.1. Modelo em árvore construído com todas as variáveis a partir da matriz binária

Numa fase inicial, foi construída uma árvore de decisão através do particionamento recursivo com base no critério do ganho associado às variáveis, sem efetuar *pruning*. Todas as variáveis foram incluídas na construção do modelo, na medida em que em cada uma das partições foi calculado o ganho associado à inclusão de cada uma delas.
Este modelo foi testado através de validação cruzada *leave one out* e foi calculada posteriormente a sua precisão, através da classificação obtida para cada uma das estirpes. O algoritmo utilizado foi construído no R. A precisão (*Acc*) e os valores preditivos foram calculados com base no número de classificações concordantes com a classificação atribuída às estirpes a partir das matriz de perda, representada na tabela 2.7.

	***C*+**	***I*+**	***N*+**
C	n_{CC}	n_{IC}	n_{NC}
I	n_{CI}	n_{II}	n_{NI}
N	n_{CN}	n_{IN}	n_{NN}

Tabela 2.7. Matriz de perda para um modelo de classificação das estirpes.

C, ***I*** e ***N*** correspondem às classificação das estirpes segundo dos dados, e ***C*+**, ***I*+** e ***N*+** à classe atribuída pelo modelo às estirpes. A precisão do classificador (*Acc*) é estimada por:

$$Acc = \frac{n_{CC} + n_{II} + n_{NN}}{n}$$

em que *n* é o número de objetos classificados. A taxa de erros da classificação é igual a 1-*Acc*, pelo que é equivalente utilizar a precisão ou a taxa de erros para avaliar o desempenho do modelo.

O desempenho do modelo foi avaliado também através da estimação dos **valores preditivos** para cada classe de estirpes (colonizadoras, invasivas e neutras). Sendo o objetivo principal a identificação de genes associados ao potencial invasivo, o **valor preditivo para as estirpes invasivas** (***VPI***) é um índice a ser maximizado, pois permite avaliar o desempenho do modelo ao classificar uma estirpe como invasiva.

O valor preditivo de um teste define-se como a probabilidade de, dado um determinado resultado (geralmente positivo ou negativo), este corresponder à categoria a que o objeto testado pertence. A estimação dos valores preditivos é normalmente utilizada em testes de diagnóstico a fim de avaliar a qualidade de resposta dos testes, e os objetos testados correspondem às amostras biológicas utilizadas.

Neste caso os resultados são os desfechos possíveis ***C***, ***I***, ou ***N***), e o *VPI* é a probabilidade de uma estirpe de *Streptococcus pneumoniae* ser invasiva sabendo que foi classificada como invasiva.

$$VPI = \frac{n_{II}}{n_{CI} + n_{II} + n_{NI}}$$

2.3.3.2. Construção de modelos em árvore com variáveis pré-selecionadas

A maioria das variáveis (genes) não é informativa, na medida em que não tem qualquer associação com a natureza das estirpes. Por outro lado, existem muitos genes que se encontram correlacionados no que diz respeito ao seu padrão de presença/ausência (figura 2.5). Para que fosse possível selecionar genes informativos, e simultaneamente não correlacionados para a construção do modelo, tornou-se necessário formar grupos de genes baseados na semelhança em termos da sua presença nas estirpes, e em cada um deles selecionar um gene representativo. Para a formação dos grupos foi utilizada classificação hierárquica com base em várias medidas de dissemelhança, e em cada um dos grupos foi escolhido, como gene representativo, o que tem associado um maior ganho (com base nos cálculos mencionados no início de 2.3.3). Para a construção dos modelos foram utilizadas a biblioteca RPART e a função *hclust* do R (Apêndice 2). Foram construídos modelos utilizando, como matriz de dados, a matriz indicadora de presença, a matriz de probabilidades de pertença, e a matriz de *scores* das Componentes Principais obtidas a partir das variáveis, para comparar o desempenho de cada um deles.

- **Modelos construídos a partir da matriz de probabilidades (sem transformação e transformada em matriz binária)**

  O algoritmo da classificação obedeceu aos seguintes passos:

1. **Classificação hierárquica dos genes com base na sua presença em cada uma das estirpes, pelo método de Ligação Completa, através da função *hclust* do R;**

Medidas utilizadas: Coeficiente de concordância simples (a partir da matriz indicadora de presença), distância euclideana e distância correlação (a partir da matriz de probabilidades de presença).

2. **Partição do dendograma obtido em grupos;**

Foram feitas partições de 2 a 72 grupos, para que fosse possível avaliar o desempenho resultante da utilização de conjuntos de genes de várias dimensões e comparar a precisão e os valores preditivos de cada um dos classificadores correspondentes.

3. **Seleção do gene com maior ganho em cada um dos grupos;**

Em cada um dos grupos foi selecionado o gene com maior ganho à partida, como gene representante do grupo.

4. **Construção dos modelos através das funções *rpart* e *rpart.control*;**

Na função *rpart* foi selecionado o critério "information", que corresponde ao ganho *(information gain*). A função *rpart.control* permite definir o número de observações máximas permitidas em cada nó terminal, que foi fixado em 1.

Foi efetuada classificação com base nos dados da matriz indicadora de presença (em que só é possível uma partição para cada variável) e com base na matriz de probabilidades de pertença. Neste último caso, uma vez que as variáveis são contínuas, o algoritmo avalia não só a variável cuja inclusão leva a um maior ganho, mas também o ponto de charneira que leva a uma otimização desse ganho. Por esse motivo, cada variável pode ser particionada em mais do que um ponto.

5. **Avaliação do desempenho do modelo através de validação cruzada *leave one out.***

Na prática este passo foi efetuado em simultâneo com a construção dos modelos, para assim serem obtidas estimativas da precisão e valores preditivos com o máximo rigor possível: cada estirpe foi classificada com base no modelo construído a partir dos dados referentes às restantes 71 estirpes, de acordo com o procedimento da validação *leave one out*, e este processo foi repetido para todos os grupos de genes. Foram obtidas as estimativas da precisão e dos valores preditivos para cada um dos modelos, e dentro destes, para cada conjunto de genes incluídos.

- **Modelos construídos com base em Componentes Principais**

Breve explicação da Análise de Componentes Principais

A Análise de Componentes Principais é um processo que permite detetar associações entre variáveis, através das suas projeções em diferentes direções, quando representadas num plano bidimensional. As direções em que as variáveis são projetadas definem novos eixos, a partir dos quais podem ser representadas através de combinações lineares. Uma vez que o objetivo

desta metodologia é a redução da dimensionalidade dos dados, as direções das projeções que se procuram são aquelas que resultam em combinações lineares que possuam a maior variabilidade dos dados, não estando correlacionadas entre si. A vantagem desta transformação é a possibilidade de um número reduzido componentes principais (desejavelmente as duas primeiras) ser suficiente para reunir grande parte da informação fornecida pela totalidade dos dados.

A análise de Componentes Principais gera tantas combinações lineares quantas as variáveis existentes. Caso as primeiras componentes principais detenham uma parte significativa da variabilidade contida nos dados, é atingido o objetivo da redução da sua dimensionalidade.

Seja X uma matriz de dados com $n= 1,...i$ objetos e $p=1,..,j$ variáveis, que constitui uma matriz de dados centrados $(x_{ij}-\overline{x_j})$.

$$a^T x=\sum_{j=1}^{n} a_j x_j \ ,$$

a é o vetor $p\times 1$ dos pesos da projeção das variáveis que resulta na maior variância quando os dados X são projetados num plano, ou seja, é o vetor que origina a rotação dos eixos na direção em que a variabilidade de X é maior. O conjunto das projeções de todos os vetores da matriz X é dado por Xa.

A variância associada a esta projeção é dada por:

$$\sigma_a^{\,2}=a^T Va\,,$$

em que $V=X^T X$, que é a matriz de covariâncias dos dados. Assim, a variância da projeção expressa-se em função de a, que é o vetor que minimiza a variância, e V. Os vetores a são denominados **vetores próprios**. Assim, pretende-se maximizar $\sigma_a^{\,2}$, o que é possível derivando a expressão $a^T Va$ e igualando a 0. Esta equação conduz à obtenção dos **valores próprios** da matriz V, que são medidas da variância associada a cada componente principal. A soma dos valores próprios é igual à variância total dos dados projetados:

$$\sigma_a^{\,2}=\sum_{j=1}^{p} a\lambda_j$$

A variância explicada pelas primeiras k componentes principais é obtida somando os seus valores próprios a partir de $\sum_{j=1}^{k}\lambda_j$, e a sua proporção relativamente à totalidade das componentes principais corresponde a :

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

em que o denominador corresponde ao traço da matriz de covariâncias (soma da variância total.)

As primeiras componentes principais correspondem aos vetores próprios associados aos maiores valores próprios da matriz de covariâncias *V*.

As componentes principais, como transformações lineares a partir das variáveis em estudo, podem ser representados por

$$Y_j = a_{1j} X_1 + a_{2j} X_2 + \ldots + a_{pj} X_p$$

O objetivo é obter novas variáveis não correlacionadas, transformadas a partir dos vetores de observações $x_1, x_2, \ldots, x_p$ das variáveis da matriz *X*. A nova matriz de dados *Y*, de dimensão $n \times p$ é obtida através dos vetores próprios:

$$y_{ij} = a_{1j} x_{i1} + a_{2j} x_{i2} + \ldots + a_{pj} x_{ip}$$

y_{ij} representa o *score* do *i*-ésimo objeto para a *j*-ésima componente principal. Os *scores* constituem assim, as variáveis transformadas, resultantes de combinações lineares da matriz de dados original *X*, podendo ser trabalhadas como novas variáveis.

No caso em que as variáveis que compoem *X* têm diferentes escalas e/ou são de naturezas diferentes, devem ser standartizadas, ou deve ser utilizada a matriz de correlações na análise em vez da matriz de covariâncias. No presente estudo, em que as variáveis têm a mesma natureza e escala (representam probabilidades de presença de genes) não foi necessária essa transformação.

Resumo do algoritmo de classificação através de Componentes Principais

O algoritmo assemelha-se aos anteriores, mas desta vez foi utilizada a matriz de *scores* como matriz de dados:

1. Classificação hierárquica dos genes com base na sua presença em cada uma das estirpes, através do método de Ligação Completa, através da função *hclust* do R. Para

a classificação hierárquica foram utilizados o coeficiente de concordância simples para a matriz indicadora de presença e a distância euclideana e a distância correlação para a matriz de probabilidades.

2. Partição do dendograma obtido em grupos;
3. Seleção do gene com maior ganho em cada um dos grupos;
4. Construção dos modelos através das funções *rpart* e *rpart.control*;

A matriz de dados de entrada foi a matriz de probabilidades de pertença, mas a classificação foi efetuada a partir da matriz de *scores* resultante da análise de Componentes Principais: em cada um dos ciclos em que é incluída uma nova variável (genes provenientes dos grupos da classificação hierárquica), é feita a análise de Componentes Principais, e a respetiva matriz de *scores* constitui a matriz de dados do classificador. Como as variáveis são contínuas, o seu particionamento é feito de acordo com a avaliação do ponto de charneira que leva ao maior ganho no sistema e as variáveis podem ser particionadas em mais do que um ponto (à semelhança do que acontece com a classificação a partir da matriz de probabilidades). O modelo em árvore é constituído, desta vez, pelas componentes principais, o que significa que as variáveis da classificação resultam da combinação linear de todas as variáveis da matriz de dados de entrada.

5. **Avaliação do desempenho do modelo através de validação cruzada *leave one out*.**

A validação cruzada foi efetuada através do mesmo procedimento aplicado à classificação através da matriz de dados, com a diferença de que os cálculos da precisão e valores preditivos são feitos com base na matriz de *scores* das Componentes Principais obtidas a partir genes representativos, e não das suas probabilidades de pertença.

3. Resultados e Discussão

3.1. Resultados da Classificação Hierárquica das estirpes

A análise exploratória e a aplicação dos vários métodos de classificação foram feitos através do *software* R, versão 2.10.1, utilizando algumas bibliotecas disponíveis ou funções criadas segundo as necessidades do estudo.

A classificação não supervisionada das estirpes foi efetuada através do método de ligação completa, pela sua propriedade de formar *clusters* esféricos e bem individualizados (2.3.1.3). Os dendogramas resultantes foram particionados em três grupos, com o objetivo de averiguar se os agrupamentos formados coicidiam com os grupos formados pelas estirpes colonizadoras, invasivas e neutras.
Foram efetuadas duas classificação utilizando as seguintes medidas de dissemelhança/semelhança:

- ***Distância euclideana*** calculada a partir das probabilidades de pertença de cada gene a cada estirpe. Uma vez que as probabilidades de pertença (*pij*) são valores de natureza contínua com a mesma amplitude de variação e de natureza semelhante, não foi necessário efetuar standartização. As distâncias euclideanas foram calculadas através da função *dist* do R.

- ***Coeficiente de concordância simples*** (*CCS*) - calculado a partir da matriz indicadora de presença (2.2.2). Em cada par de objetos *i* e *j* a presença ou a ausência de cada gene nos dois objetos em simultâneo deve ser igualmente valorizada, visto que se está a comparar a composição genética das estirpes. O *CCS* foi calculado através de uma função criada para o efeito (Apêndice 1 – 1.1), e transformado em medida de distância através da conversão 1-*CCS*.

Os dendogramas respeitantes às duas classificações são apresentados nas figuras 3.1 e 3.2.

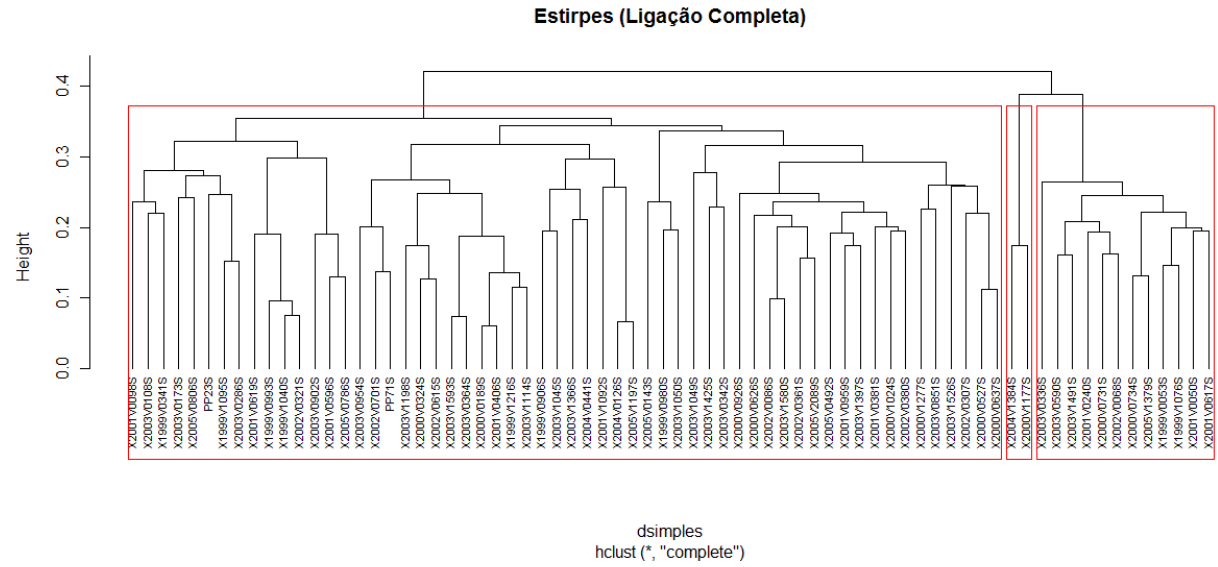


Figura 3.1. Dendograma das estirpes utilizando o ***coeficiente de concordância simples****, pelo método de ligação completa.*

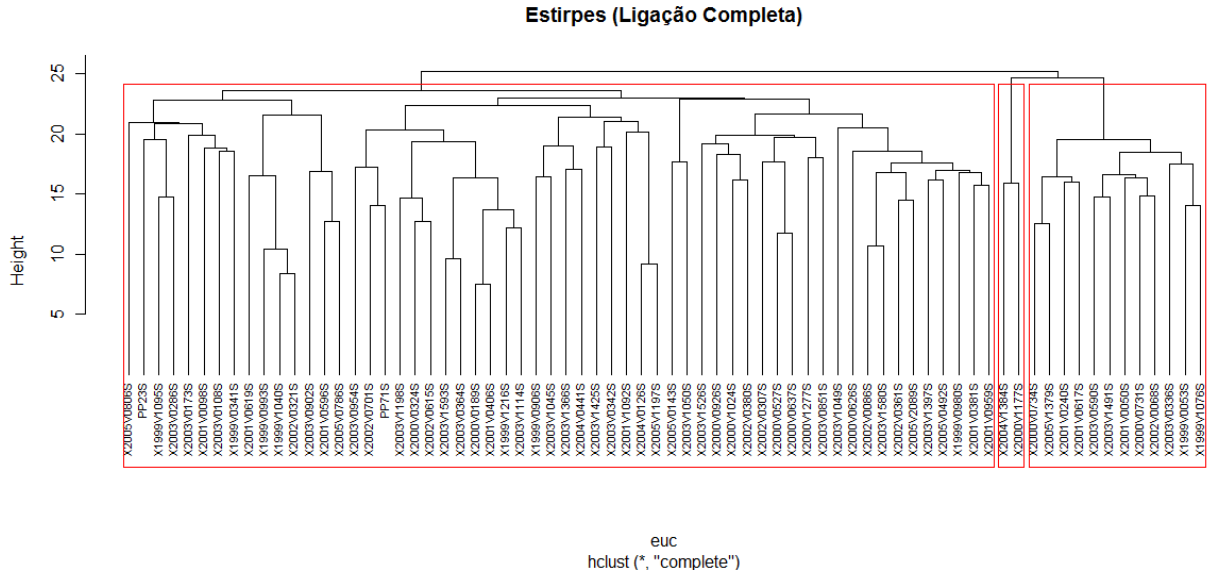


Figura 3.2. Dendograma das estirpes utilizando ***distância euclideana****, pelo método de ligação completa.*

Os resultados mostram que as duas medidas utilizadas deram origem aos mesmos agrupamentos. Para verificar até que ponto existe concordância entre os grupos formados e a classificação das estirpes, foi feita a sua comparação através de uma imagem representativa das concordâncias entre os grupos, e calculados os respetivos coeficientes de concordância simples (figura 3.3 e tabela 3.1).

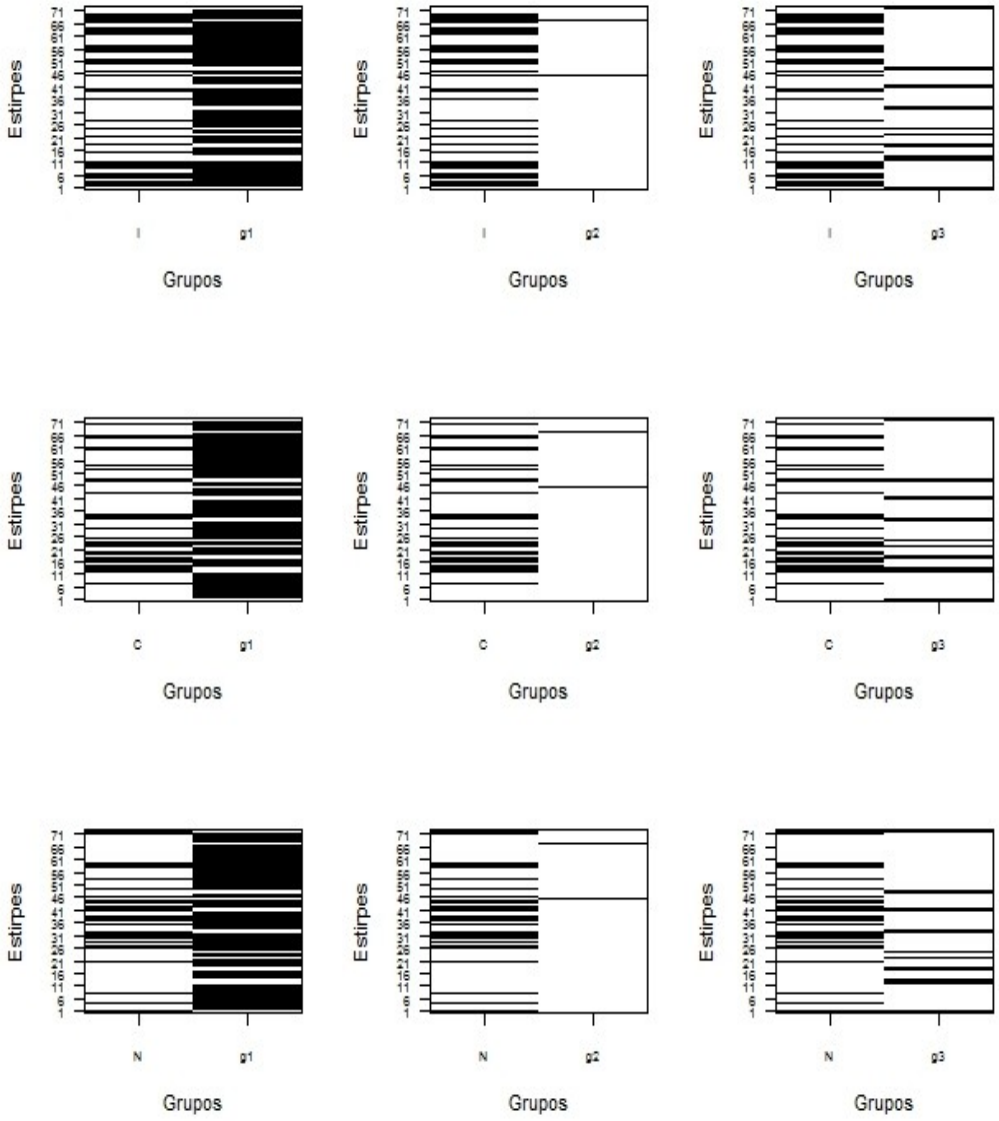


Figura 3.3 Imagem representativa dos grupos de estirpes obtidos pela classificação hierárquica , e comparação com os grupos aos quais pertencem. As zonas a negro representam estirpes pertencentes ao grupo e as zonas a branco as que não foram incluídas no mesmo. As seis imagens correspondem às combinações possíveis entre os grupos pré-classificados e aqueles que foram obtidos na classificação hierárquica.

	g1	*g2*	*g3*
I	0,49	0,63	0,49
C	0,32	0,68	0,71
N	0,39	0,67	0,64

Tabela 3.1. Coeficientes de concordância simples entre os grupos respeitantes à classificação das estirpes (I, C e N) e os grupos originados pela classificação hierárquica (g1, g2 e g3).

Na figura 3.3 e não é possível visualizar semelhanças entre os entre os grupos obtidos através da classificação hierárquica e a classificação das estirpes segundo o seu grau de invasibilidade. Os coeficientes de concordância simples calculados para as diferentes combinações entre a classificação das estirpes e os grupos formados também não são conclusivos, não permitindo identificar claramente semelhanças entre os grupos, o que leva a concluir que este não é um método eficaz na procura de associações entre o conteúdo genético das estirpes e o seu potencial invasivo.

3.2. Resultados da Classificação Supervisionada através de Modelos em Árvore

3.2.1. Modelo em árvore construído com todas as variáveis do genoma acessório

Numa primeira abordagem foi construído um classificador, através de funções construídas no R, utilizando todas as variáveis disponíveis (1775 genes do *accessory genome*). As funções utilizadas permitiram calcular a entropia de cada sistema, a entropia condicional a cada variável, a classificação obtida para as estirpes ao longo da árvore (em cada ciclo de partição), e a classificação de cada estirpe obtida no processo de validação cruzada *leave one out*. A avaliação da capacidade preditiva foi efetuada com base nos resultados da validação cruzada através do cálculo da precisão e dos valores preditivos para as estirpes colonizadoras, invasivas e neutras (*VPC*, *VPI* e *VPN* respetivamente).

	C+	***I***+	***N***+
C	6	5	10
I	5	18	6
N	7	10	5

Tabela 3.2. Resultado da classificação das estirpes através de validação cruzada leave one out. As linhas referem-se à classificação conhecida das estirpes e as colunas à categoria atribuída pelo modelo em árvore na validação cruzada.

Precisão	0,4
VPC	0,33
VPI	0,55
VPN	0,24

Tabela 3.3. Precisão (*Acc*) e valores preditivos do classificador, em que *VPC*, *VPI* e *VPN* correspondem aos valores preditivos para as estirpes colonizadoras, invasivas e neutras, respetivamente.

Os resultados não foram satisfatórios, tendo-se obtido uma precisão na classificação inferior a 50%. Os valores preditivos correspondentes às diferentes classes de estirpes demonstram que o classificador não pode ser utilizado como modelo preditivo. Apenas o valor preditivo obtido das estirpes invasivas (*VPI*=0,55) é superior à sua frequência relativa no conjunto de dados (0,40). Os valores preditivos correspondentes às estirpes colonizadoras e neutras não diferem substancialmente das suas frequências relativas (0,29 e 0,30 respetivamente), o que leva a concluir que a sua classificação, segundo este modelo, pode ter resultado de variáveis não informativas e/ou em partições dos ramos terminais em que o número de observações é demasiado pequeno para representar associações entre a variável e o tipo de desfecho.

3.2.2. Modelos em árvore construídos com variáveis pré-selecionadas

Partindo de três critérios de seleção dos genes a serem utilizados na construção dos classificadores (classificação hierárquica utilizando três medidas de distância diferentes), foram construídos modelos de classificação a partir dos dados em matriz binária, matriz de probabilidades e matriz de *scores* das Componentes Principais (métodos descritos em 2.3.3.2.).

Assim, dispõe-se de nove cenários diferentes a serem comparados, a fim de averiguar as vantagens e desvantagens de cada método e comparar a precisão e o *VPI* em cada um dos casos. De seguida são apresentados os resultados obtidos a partir dos genes selecionados segundo os três critérios, conforme se encontra descrito em 2.3.3.2 (ponto 1).

- **Coeficiente de Concordância Simples como medida utilizada no cálculo da dissemelhança entre os genes**

Os gráficos das figuras 3.4, 3.5 e 3.6 descrevem a precisão e os valores preditivos para as estirpes colonizadoras, invasivas e neutras (*VPC*, *VPI* e *VPN* respetivamente) obtidos no caso em que são incluídas, na classificação, p=2,...,72 variáveis (genes), selecionadas através da classificação hierárquica e posterior formação de grupos, dos quais é selecionado o que apresenta maior ganho. Os índices foram estimados a partir dos resultados da validação cruzada *leave one out*.

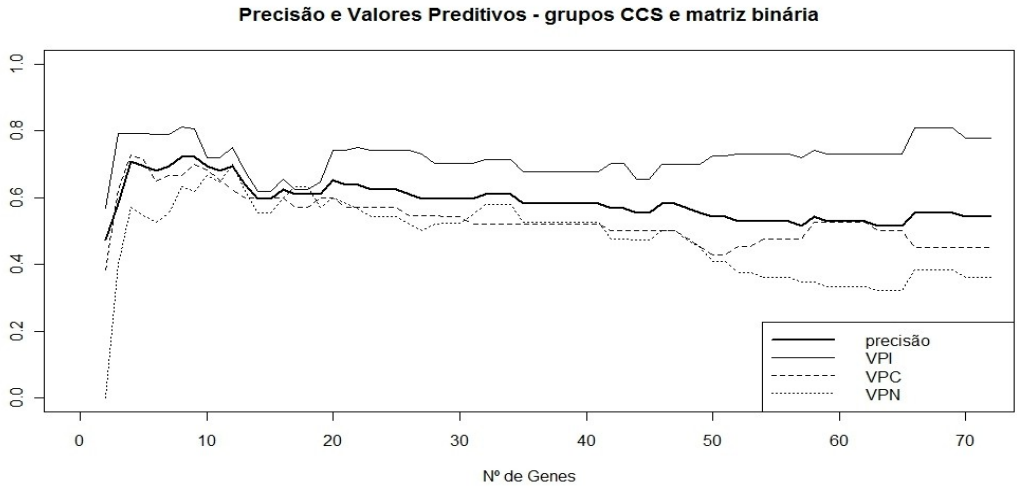


Figura 3.4. Resultados da classificação a partir da matriz indicadora de presença.

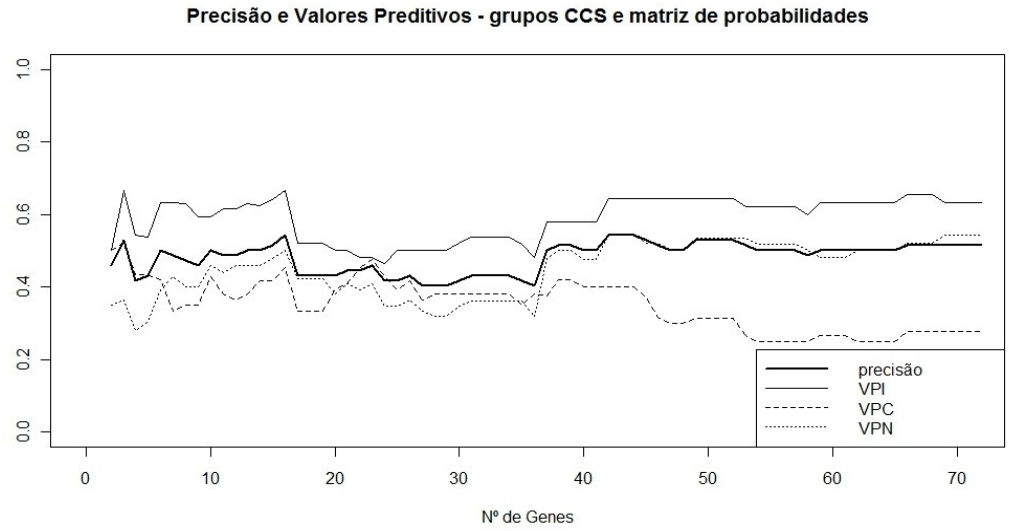


Figura 3.5. Resultados da classificação a partir da matriz de probabilidades.

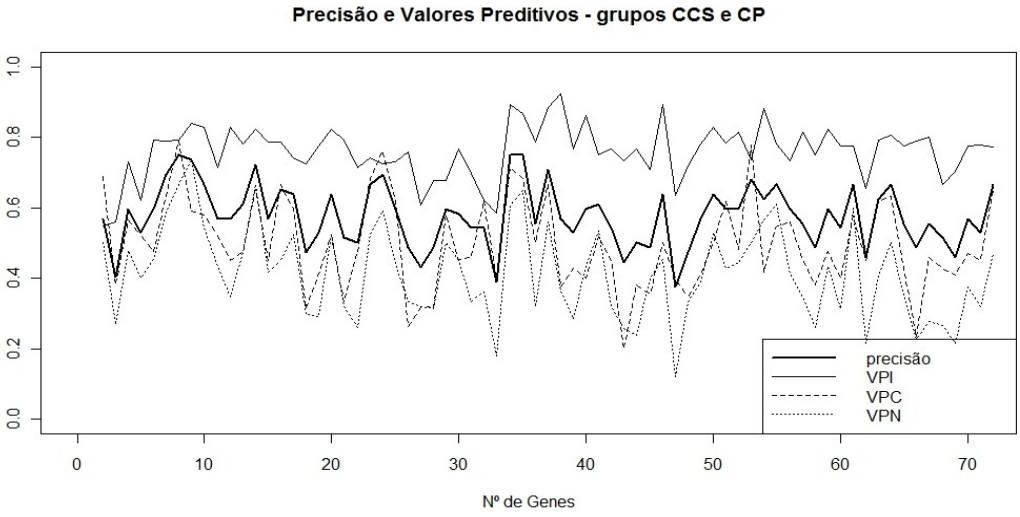


Figura 3.6. Resultados da classificação a partir da matriz de scores *das* CP.

Os gráficos mostram que, entre os valores preditivos, o *VPI* é o que atinge níveis mais elevados, seguido do *VPC* e do *VPN*, o que sugere que as variáveis utilizadas na classificação têm uma associação mais forte com o potencial invasivo do que com o caráter colonizador ou neutro das estirpes. Os baixos *VPN* observados poderão dever-se ao facto de estas estirpes, não sendo claramente invasivas ou colonizadoras, apresentarem um caráter intermédio (no que diz respeito à sua composição genética) que as torma mais difíceis de classificar.

A precisão (*Acc*) parece refletir e acompanhar as variações observadas nos valores preditivos

ao longo do gráfico, sendo uma medida global da capacidade preditiva dos modelos.

A observação das figuras 3.4 e 3.5 leva a concluir também que não são necessárias muitas variáveis para atingir os níveis de precisão e valores preditivos máximos. A classificação com base em componentes principais constitui a exceção.

A tabela 3.4 apresenta os dados referentes aos modelos em que é maximizada a precisão e o *VPI*. A primeira coluna indica qual a matriz de variáveis utilizada na classificação (presença/ausência, probabilidade de presença ou *scores* das componentes principais construídas a partir da matriz de probabilidades de ausência). A segunda coluna indica o valor máximo de precisão obtido e a terceira coluna apresenta as variáveis (genes) utilizadas na construção desse modelo (entre parêntesis está indicado o número de grupos dos quais foram extraídos os genes representativos que constituíram a matriz de dados de entrada do modelo). A quarta e quinta colunas apresentam a mesma informação, mas desta vez referente ao modelo em que se atingiu o *VPI* máximo.

Na última linha, referente à classificação com base em componentes principais, as colunas referentes aos genes utilizados na classificação encontram-se divididas em duas, a primeira onde são indicadas as CP utilizadas na construção dos modelos, e a segunda os genes a partir dos quais estas foram construídas.

Matriz de dados	***Precisão máxima***			***VPI máximo***		
	Valor	***Genes utilizados na classificação***		***Valor***	***Genes utilizados na classificação***	
BINÁRIA	0,72	1058 **1223 2118 2593 2605** 3055 3186 **3193** *(8 de 8 grupos)*		0,81	**1058** 1223 **2118** 2593 2605 3055 3186, 3193 *(8 de 8 grupos)*	
PROBABILIDADES	0,54	185, 923, **1223**, **2118**, **2593**, **2605**, 2777, **3193**, 3307 3393 *(10 de 16 grupos)*		0,67	**1058 2118** *(2 de 2 grupos)*	
COMPONENTES PRINCIPAIS	0,75	***CP***	***Genes***	0,92	***CP***	***Genes***
		1 2 3 4 5 6 8 *(7 CP)*	1058 **1223 2118 2593 2605** 3055 3186 **3193** *(Combinações lineares de 8 genes)*		1 2 6 7 9 14 19 20 23 37 *(10 CP)*	86 185 223 754 755 923 972 **1058** 1223 1230 1363 1565 1704 1712 1938 1977 1978 **2118** 2161 2591 2593 2604 2605 2635 2777 2944 2971 3055 3132 3186 3193 3307 3376 3393 3495 3516 3522 3565 *(Combinações lineares de 38 genes)*

Tabela 3.4. Resumo dos dados referentes aos modelos de classificação com precisão e VPI máximos (com o número mínimo de variáveis). Seleção dos genes representativos: classificação hierárquica com base em coeficiente de concordância simples. A *bold* estão indicados os genes comuns aos três métodos de classificação.

Os modelos obtidos a partir da matriz binária apresentam melhores resultados de precisão e *VPI* em comparação com os que foram construídos a partir da matriz de probabilidades (0,72 e 0,81 respetivamente), o que leva a crer que as variáveis, expressas em termos de

presença/ausência em vez de probabilidades de presença, têm melhor capacidade discriminativa ao serem utilizadas num classificador em árvore. Neste caso existe apenas um ponto de charneira possível, enquanto que as variáveis expressas em probabilidades, sendo contínuas, têm um número infinito de pontos charneira possíveis, podendo ser particionadas em mais do que um ponto. Os resultados levam assim a concluir que a partição segundo o critério referido em 2.2.1 constitui o valor de charneira que leva à otimização de um modelo em árvore que se pretenda que seja preditivo.

Os modelos obtidos a partir de componentes principais, comparativamente com os anteriores, apresentam globalmente melhores resultados de precisão e *VPI* (0,75 e 0,92), talvez devido à capacidade que as componentes principais têm de refletir a variabilidade de muitos genes em combinações lineares. No entanto a figura 3.6 revela grandes oscilações dos valores preditivos e precisão à medida que são incluídas novas variáveis, que transmite uma ideia de instabilidade, ou falta de robustez, deste método.

As figuras 3.7, 3.8 e 3.9 representam os modelos em árvore construídos com base nos resultados em que a precisão e o *VPI* são mais elevados. A primeira árvore corresponde ao modelo construído a partir da matriz binária que apresenta simultaneamente valores máximos de precisão e *VPI*, e as duas seguintes aos modelos construídos com base em *CP* com precisão e *VPI* máximos, respetivamente. Estes modelos foram construídos a partir dos dados de todas as estirpes.

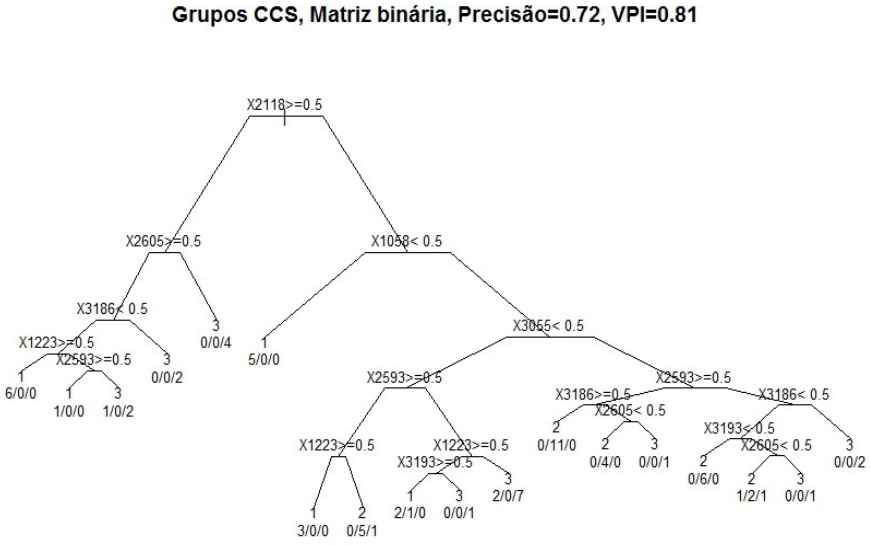


Figura 3.7. A variável que preside a cada partição (número do gene) e o seu ponto de charneira são representados em cada nó da árvore. O critério de partição representado refere-se ao ramo do lado esquerdo, ou seja, as estirpes para as quais, por exemplo, X2118 ≥ 0,5 (gene presente) são alocadas no ramo esquerdo da árvore, e aquelas para os quais X2118<0,5 (gene ausente), no ramo do lado direito. A classificação final das estirpes encontra-se representada por números, em que 1- Colonizadora, 2- Invasiva e 3 – Neutra. O número de estirpes classificadas em cada ramo terminal encontra-se representado pela mesma ordem.

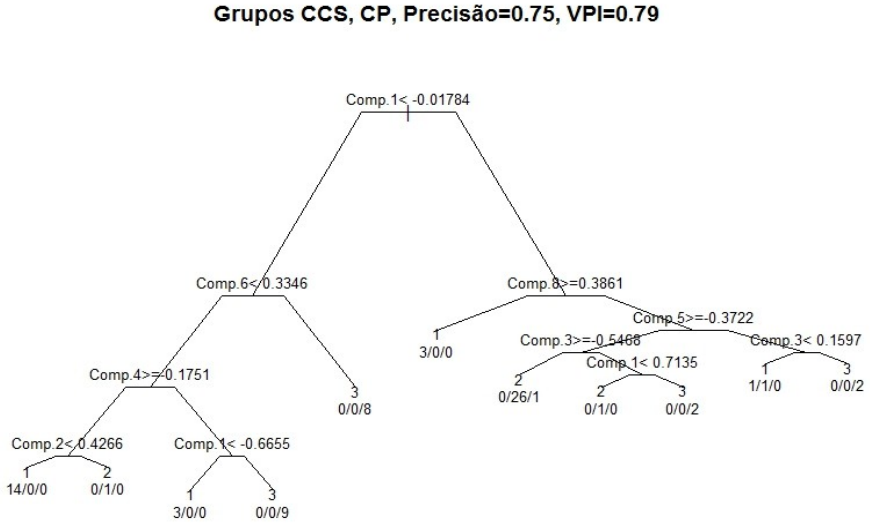


*Figura 3.8. Modelo em árvore construído a partir das CP das variáveis, em que é maximizada a precisão. Neste caso as variáveis (*scores *das CP) são contínuas, e os pontos de charneira que condicionam a partição em cada ramo são calculados por forma a que cada partição resulte na máxima redução da entropia.*

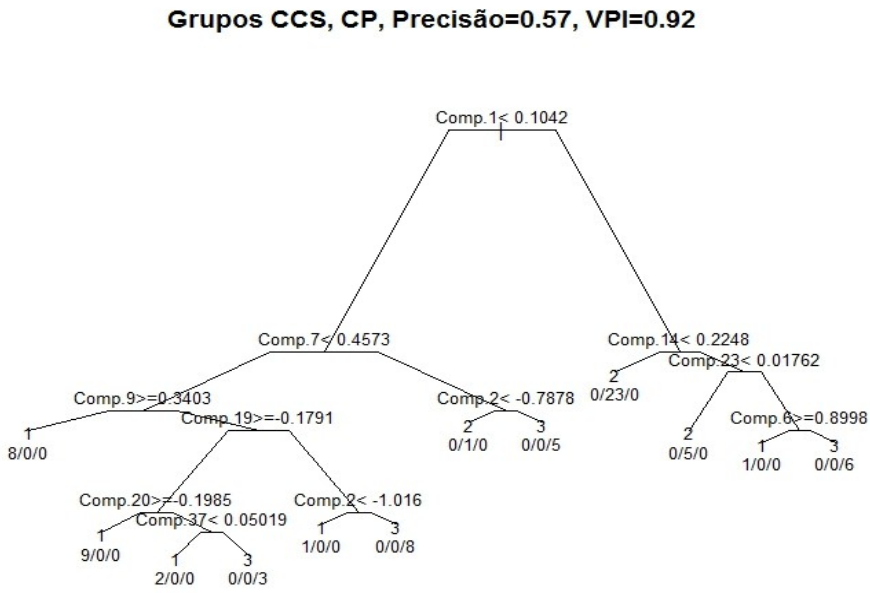


Figura 3.9. Modelo em árvore construído a partir das CP das variáveis, em que é maximizado o VPI.

O modelo da figura 3.7 (CCS, matriz binária) constitui aquele em que a precisão e o VPI são simultaneamente maximizados. Os modelos construídos com base nas CP demonstram ser possível obter valores ainda mais elevados, no entanto, tal como foi referido anteriormente, o método parece não ser robusto, uma vez que a introdução de variáveis provoca grandes oscilações nos valores preditivos e precisão. De facto, ao analisar com mais detalhe as componentes principais obtidas, verifica-se que, ao contrário do que é esperado, as primeiras componentes principais não conseguem reunir a maioria da variabilidade dos dados.

Tomando como exemplo o modelo da figura 3.9 (CP construídas a partir de 38 variáveis) o gráfico dos *scores* correspondentes às duas primeiras CP revela não existir uma direção preferencial na distribuição dos pontos, o que leva a concluir que estas não retêm a maior parte da variabilidade dos dados. O *screeplot* correspondente (gráfico das variâncias retidas pelas componentes principais), confirma esta observação (figuras 3.10 e 3.11). Estas observações são uma explicação possível para as oscilações observadas no gráfico da figura 3.6.

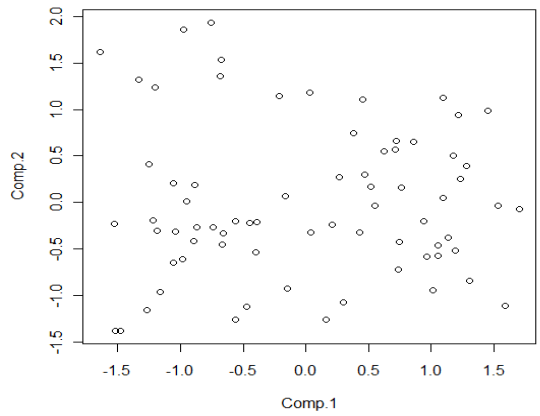


Figura 3.10. Gráfico dos scores *das CP 1 e 2.*

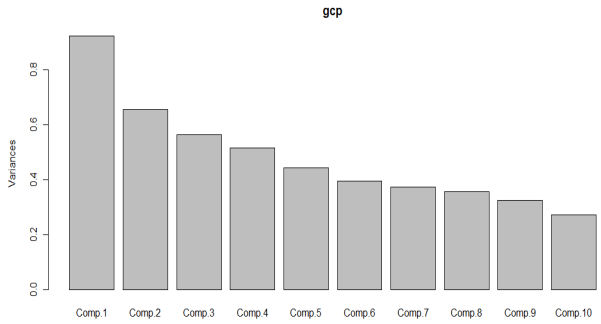


Figura 3.11. Screeplot *das 10 primeiras componentes principais.*

O facto de os genes a partir dos quais foram construídas as componentes principais terem sido selecionados de forma a terem entre si o mínimo em comum (selecionados de grupos diferentes), faz com que a correlação existente entre si seja reduzida. A tabela 3.5 representa a matriz de correlações entre os 8 genes utilizados no modelo construído com base em componentes principais em que a precisão é máxima, onde se pode constatar que a correlação é fraca, ou quase nula. Esta observação poderá ser a razão de as primeiras componentes principais não explicarem a maior parte da variabilidade dos dados, como seria desejável.

	1223	***2605***	***2118***	***2593***	***1058***	***3055***	***3193***	***3186***
1223	1,00	0,27	0,33	-0,04	-0,11	-0,15	0,07	-0,13
2605	0,27	1,00	0,28	-0,06	-0,05	-0,24	0,14	0,08
2118	0,33	0,28	1,00	-0,22	0,00	-0,17	0,17	-0,16
2593	-0,04	-0,06	-0,22	1,00	0,09	0,17	0,02	0,36
1058	-0,11	-0,05	0,00	0,09	1,00	-0,01	-0,11	0,11
3055	-0,15	-0,24	-0,17	0,17	-0,01	1,00	-0,34	0,15
3193	0,07	0,14	0,17	0,02	-0,11	-0,34	1,00	-0,24
3186	-0,13	0,08	-0,16	0,36	0,11	0,15	-0,24	1,00

Tabela 3.5. Matriz de correlações entre os 8 genes representativos de 8 grupos, utilizados no modelo com com base em componentes principais em que a precisão é máxima.

- **Distância Euclideana como medida utilizada no cálculo da dissemelhança entre os genes**

  Gráficos da precisão e valores preditivos:

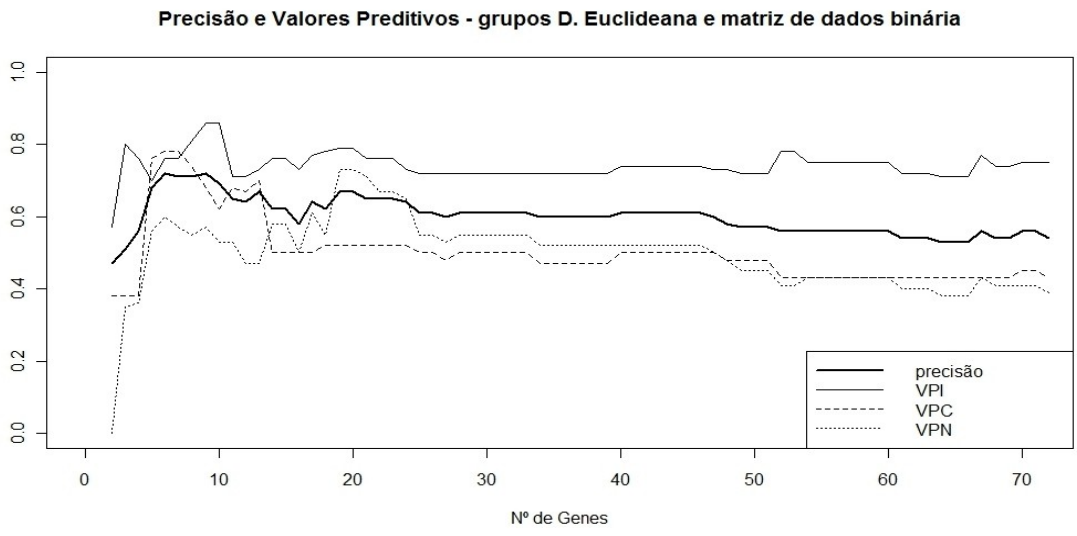


Figura 3.12. Resultados da classificação a partir da matriz indicadora de presença.

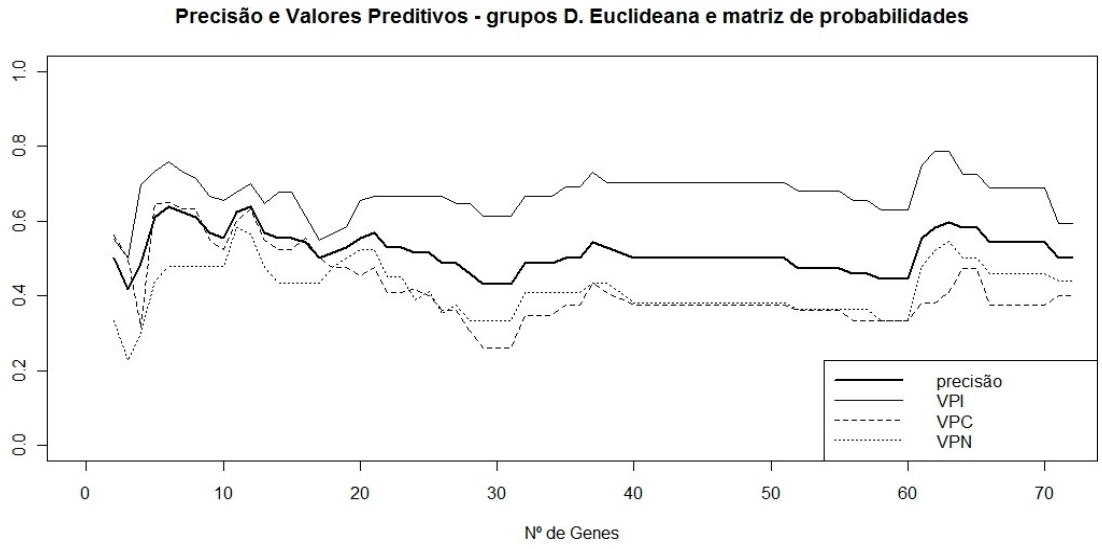


Figura 3.13.Resultados da classificação a partir da matriz de probabilidades.

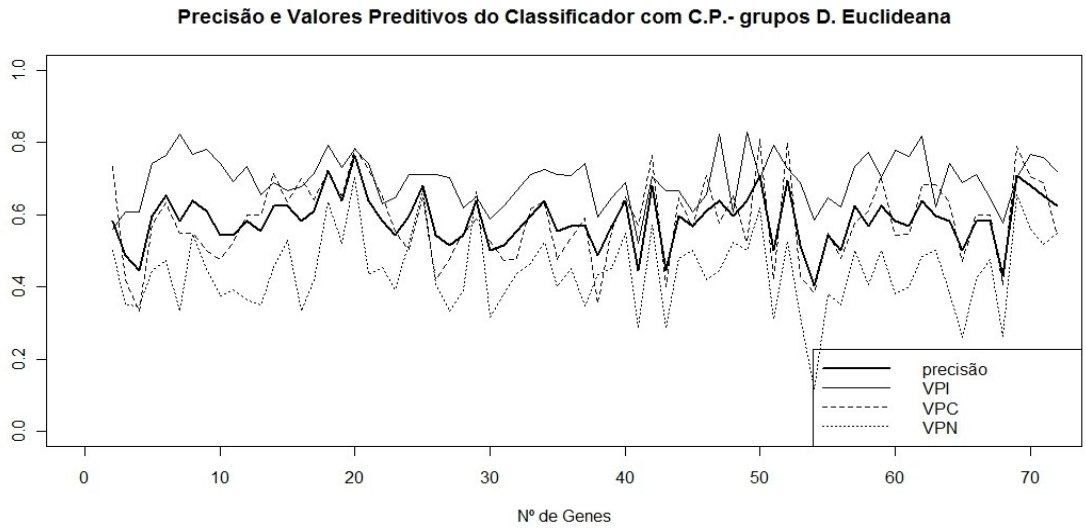


Figura 3.14. Resultados da classificação a partir da matriz de scores *das* CP.

A tabela 3.6 resume os dados correspondentes aos modelos em que são atingidos a precisão e *VPI* máximos, à semelhança da tabela 3.4.

Matriz de dados	***Precisão máxima***			***VPI máximo***		
	Valor	***Genes utilizados na classificação***		***Valor***	***Genes utilizados na classificação***	
BINÁRIA	0,72	**1058 1619 2118 2593 2605** *(5 de 6 grupos)*		0,86	1058 1619 **2118 2593 2605** 3186 3193 3283 *(8 de 9 grupos)*	
PROBABILIDADES	0,64	**1058 1619 2118 2593 2605** 3193 *(6 de 6 grupos)*		0,79	955 1062 2114 **2118** 2159 **2593 2605** 2788 2841 3283 3307 *(11 de 62 grupos)*	
COMPONENTES PRINCIPAIS	0,76	***CP***	***Genes***	0,82	***CP***	***Genes***
		1, 3, 4, 5, 14, 16 *(6 CP)*	252 363 923 **1058** 1485 **1619** 1704 1977 **2118** 2159 2163 **2593 2605** 3055 3186 3193 3283 3307 3393 3565 *(Combinações lineares de 20 genes)*		1, 3, 4, 5, 6, 7 *(6 CP)*	1058 1619 1977 **2118 2593 2605** 3193 *(Combinações lineares de 7 genes)*

Tabela 3.6. Resumo dos dados referentes aos modelos de classificação com precisão e *VPI* máximos, respetivamente (com o número mínimo de variáveis). Seleção dos genes representativos: classificação hierárquica com base em distância euclideana.

No que diz respeito à classificação a partir da matriz binária e de probabilidades, os valores de precisão e *VPI* máximos obtidos a partir deste critério de seleção dos genes (0,72 e 0,86 respetivamente) são semelhantes aos anteriores, com a diferença que, para atingir o mesmo nível de precisão foram necessárias menos variáveis (5 genes). Os modelos construídos a

partir da matriz de probabilidades apresentam mais uma vez valores inferiores aos construídos a partir da matriz binária, pelas razões expostas anteriormente. A classificação a partir de CP, neste caso, não introduz melhorias substanciais nos resultados comparativamente com classificação a partir da matriz binária, não trazendo, neste caso, qualquer vantagem.

A figura seguinte representa o modelo em árvore construído a partir da matriz binária no qual se encontram maximizados simultaneamente a precisão e o *VPI*.

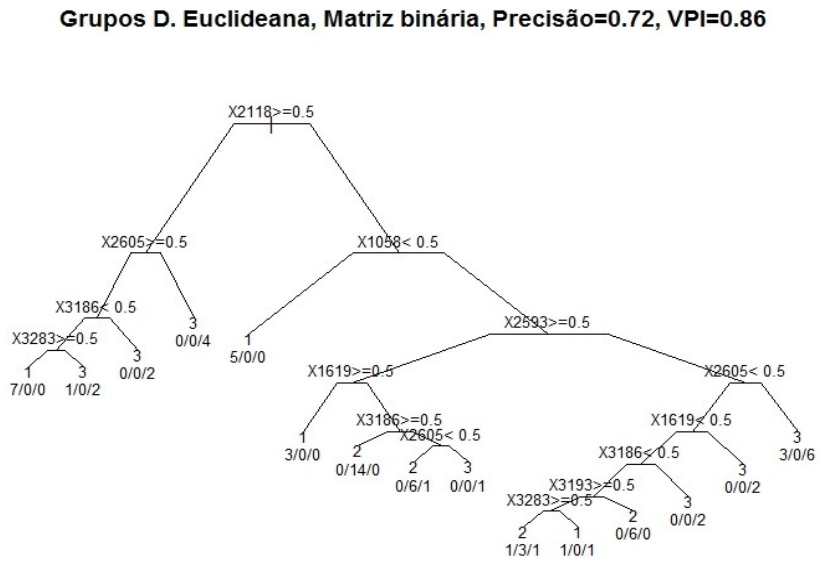


Figura 3.15. Modelo em árvore construído a partir da matriz binária no qual se encontram maximizados simultaneamente a precisão e o VPI.

Este modelo é comparável ao representado na figura 3.7, uma vez que utiliza o mesmo número de variáveis na sua construção e permite obter o mesmo grau de precisão e um *VPI* ligeiramente mais elevado (0,86).

- **Distância Correlação como medida utilizada no cálculo da dissemelhança entre os genes**

  Gráficos da precisão e valores preditivos:

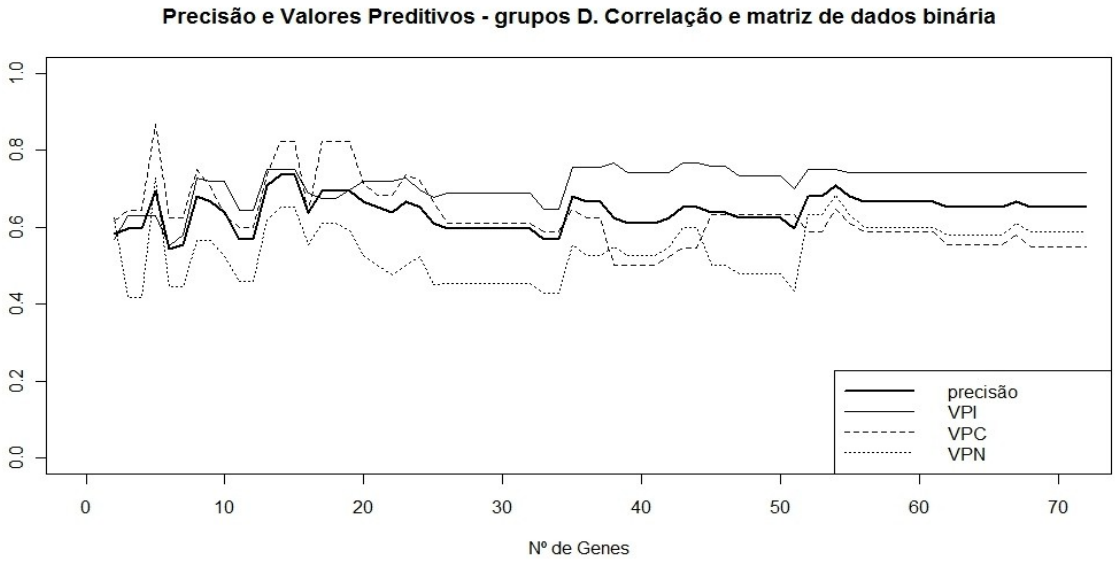


Figura 3.16. Resultados da classificação a partir da matriz binária.

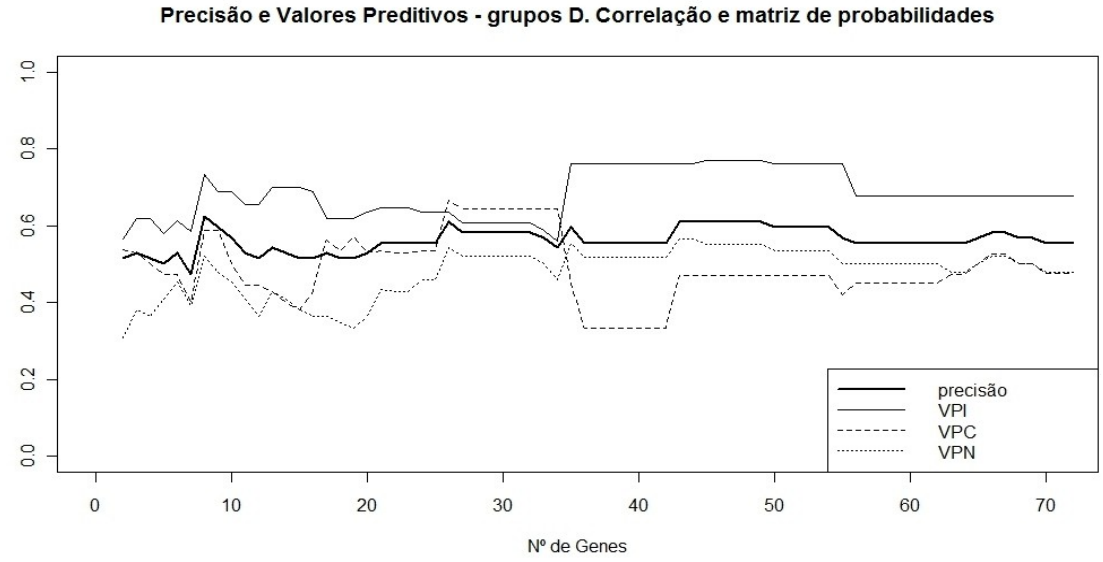
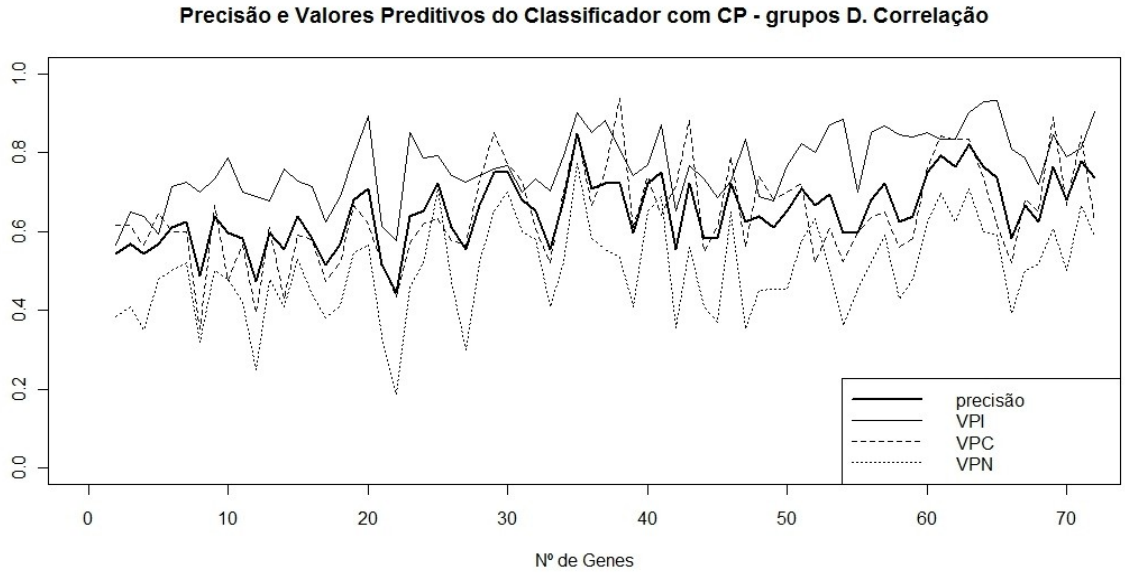


Figura 3.17. Resultados da classificação a partir da matriz de probabilidades.

Figura 3.18. Resultados da classificação a partir da matriz de scores *das* CP.



Matriz de dados	***Precisão máxima***			***VPI máximo***		
	Valor	***Genes utilizados na classificação***		***Valor***	***Genes utilizados na classificação***	
BINÁRIA	0,74	265 **1058** **1181** 1619 1620 1872 **2118** 2142 2564 **2635** 3172 **3238** *(12 de 14 grupos)*		0,77	104 754 972 1058 1181 1452 1485 1773 **2118** 2142 2469 **2593** 2600 2635 2777 3172 **3193** 3544 *(18 de 38 grupos)*	
PROBABILIDADES	0,63	754 **1058** **1181** **2118** **2635** **3238** *(6 de 8 grupos)*		0,77	30 222 265 289 1872 **2118** 2453 **2593** **3193** 3500 *(10 de 45 grupos)*	
COMPONENTES PRINCIPAIS	0,85	***CP***	***Genes***	0,93	***CP***	***Genes***
		1, 2, 4, 7, 10, 26, 35 *(7 CP)*	104 130 265 289 695 754 972 **1058** 1091 **1181** 1222 1452 1485 1619 1620 1773 1872 **2118** 2142 2400 2453 2469 2505 2564 2593 2600 **2635** 2994 3117 3132 3172 3193 **3238** 3393 3544 *(Combinações lineares de 35 genes)*		1, 2, 3, 5, 7, 31, 44, *(7 CP)*	30 104 107 130 185 222 265 289 363 535 618 695 754 832 850 871 951 972 1058 1091 1171 1181 1222 1246 1429 1452 1485 1619 1620 1773 1872 **2118** 2142 2162 2400 2453 2469 2505 2564 **2593** 2600 2621 2635 2720 2777 2917 2994 3055 3089 3117 3132 3172 3183 **3193** 3238 3248 3347 3386 3393 3487 3500 3516 3522 3544 3609 *(Combinações lineares de 65 genes)*

Tabela 3.7. Resumo dos dados referentes aos modelos de classificação com precisão e VPI máximos, respetivamente (com o número mínimo de variáveis). Seleção dos genes representativos: classificação hierárquica com base em distância correlação.

Os resultados obtidos a partir deste critério de escolha dos genes não diferem substancialmente dos anteriores quando se comparam os três métodos de classificação. Verifica-se um decréscimo nos *VPI* correspondentes à matriz binária e de probabilidades (tabela 3.7), mas, por outro lado, observa-se um acréscimo do *VPC* na matriz binária relativamente aos outros índices quando são utilizadas poucas variáveis (figura 3.16), o que constitui uma exceção relativamente aos outros modelos, e leva a colocar a hipótese de que alguns genes poderão estar mais associados com o caráter colonizador do que com o caráter invasivo das estirpes. É de notar também que, de um modo geral, parecem ser necessárias mais variáveis para atingir os mesmos níveis de precisão quando é utilizada a distância correlação na seleção dos genes. Nos modelos construídos com CP há de novo uma melhoria nos resultados, mas é necessário um número elevado de variáveis para a sua construção, e a robustez do modelo é posta em causa, mais uma vez, ao verificar que a introdução de novas variáveis leva a grandes oscilações nos valores preditivos e na precisão (figura 3.18).

3.3 Discussão

3.3.1. Classificação não supervisionada das estirpes através de agrupamento hierárquico

A classificação hierárquica das estirpes com posterior partição em três grupos revelou não ser um método eficaz na procura de associações entre o seu conteúdo genético e a classe em que se encontram classificadas. Uma causa possível para estes resultados é o enorme número de variáveis utilizadas (1775), pois grande parte delas poderá não ter qualquer relação com o potencial invasivo da espécie.

Uma explicação possível para os agrupamentos obtidos através da classificação hierárquica, é a hipótese de estes resultarem das relações filogenéticas entre as estirpes.

3.3.2. Classificação supervisionada das estirpes através de modelos em árvore

O primeiro modelo em árvore construído (3.2.1) incluiu todas as variáveis na matriz de dados de entrada, sendo selecionada, em cada passo, aquela que resultava numa maior redução de entropia em cada subsistema. A validação cruzada revelou, no entanto, que a precisão e os

valores preditivos obtidos através da aplicação deste método se encontram muito aquém do desejável (tabela 3.3). Dada a dimensão do conjunto de dados e o facto de muitos genes serem semelhantes no que diz respeito ao seu padrão de presença nas estirpes, tornou-se necessário agrupar os genes com base na sua semelhança e escolher um representante de cada grupo, para evitar a inclusão de variáveis muito semelhantes que fornecem a mesma informação (2.3.3.2). O gene representante de cada grupo foi, naturalmente, aquele cujo ganho associado à sua inclusão num modelo em árvore era maior.

Este método, testado para três medidas de dissemelhança no agrupamento dos genes, e com grupos de genes representativos de várias dimensões, resultou numa melhoria significativa da precisão e valores preditivos ao classificar as estirpes. No que diz respeito à matriz de dados de entrada (matriz indicadora de presença, de probabilidades ou de componentes principais dos genes), os modelos em árvore construídos a partir da matriz indicadora de presença revelam ser mais precisos do que os resultantes da matriz de probabilidades, em todos os casos. A diferença fundamental é o ponto de charneira que origina a partição das variáveis. Uma vez que a matriz indicadora de presença foi construída tendo em conta a aplicação do modelo NUDGE e o critério que distingue os genes de cada subpopulação (2.2.1), essa pode ser a razão para a melhor capacidade preditiva observada nos modelos construídos a partir da mesma, em comparação com os modelos construídos a partir da matriz de probabilidades.

Os modelos construídos a partir das componentes principais apresentam de uma forma geral melhores resultados em comparação com os obtidos com as variáveis sem transformação (matriz de probabilidades). Por um lado, isso revela que as componentes principais que entram nos modelos conseguem, de certa forma, reunir informação relevante acerca das variáveis; no entanto, esta metodologia revela problemas quando se observa que a introdução de novas variáveis no modelo leva a grandes oscilações nos valores preditivos e precisão, que poem em causa a robustez deste método.

A observação do gráfico das primeiras CP e o respetivo *screeplot* revelam que estas não conseguem reter a maioria da variabilidade dos dados, pelo que a introdução de qualquer variável pode levar a oscilações grandes nos resultados. A explicação para esta observação reside no facto de as variáveis terem sido escolhidas entre grupos formados com base nas suas dissemelhanças ou correlações: em cada grupo de variáveis que se assemelham mais entre si, só uma delas (gene representante do grupo) entra no modelo. Isso faz com que todas

as variáveis sejam pouco correlacionadas entre si, e as CP (que são construídas com base nas covariâncias/correlações entre as variáveis) não consigam reter uma estrutura inerente aos dados, o que pode levar a que os modelos sejam mais sensíveis à introdução de novas variáveis.

4. Conclusões

No que diz respeito às metodologias de classificação - classificação supervisionada e não-supervisionada – os resultados levam a concluir que a classificação supervisionada é a mais adequada para procurar variáveis associadas ao potencial invasivo de *Streptococcus pneumoniae*.

A classificação hierárquica das estirpes com base no seu conteúdo genético (não supervisionada) revelou não ser um método eficaz na procura dos genes mais informativos, uma vez que o número de variáveis utilizadas na classificação é excessivamente grande, e uma parte significativa das mesmas pode não estar relacionada com o potencial invasivo da espécie.

A classificação supervisionada pressupõe conhecimento prévio acerca da natureza das estirpes (que neste caso é a sua classificação a *priori*), o que constitui uma vantagem, pois permite a seleção das variáveis potencialmente mais informativas com base nesse conhecimento a partir de um conjunto de variáveis de grande dimensão. Apesar de nos modelos em árvore ser selecionada a variável mais informativa em cada uma das partições, o primeiro modelo em árvore, construído tendo como variáveis de entrada todos os genes do genoma acessório, revelou não ter capacidade preditiva. Tornou-se assim evidente a necessidade de reduzir os dados através da seleção dos genes mais informativos, que passaram a ser as variáveis de entrada para os modelos. Para que não fossem escolhidos genes muito semelhantes entre si (porque fornecem a mesma informação), foram escolhidos genes representativos de grupos formados a partir da sua classificação hierárquica com base na sua semelhança no que diz respeito à sua presença em cada uma das estirpes. Através deste método foram obtidos modelos com capacidade preditiva bastante mais elevada, dos quais se destacam os que foram construídos a partir da matriz indicadora de presença. Os modelos construídos a partir da matriz de probabilidades têm de uma forma geral valores preditivos e precisão inferiores, e os modelos com base em componentes principais, apesar da sua elevada capacidade preditiva, revelaram a desvantagem descrita em 3.3.2.

Os gráficos da precisão e valores preditivos revelam também que não é necessário um número muito elevado de genes para a obtenção de um modelo com a capacidade preditiva máxima (5 a 8 genes foram suficientes – tabelas 3.4 e 3.6).

De entre as medidas de precisão avaliadas, o *VPI* foi quase sempre mais elevado do que os restantes, sugerindo que os genes utilizados na classificação estão mais relacionados com o potencial invasivo (através da sua presença ou ausência) do que com o caráter colonizador ou neutro das estirpes. As restantes estirpes parecem ser mais difíceis de classificar, fazendo com que a precisão dos modelos seja geralmente mais baixa quando comparada com o *VPI*.

Os modelos em árvore construídos a partir de variáveis pré-selecionadas, expressas através da indicação de presença em cada uma das estirpes revelaram, neste estudo, serem o método mais eficaz para predizer a capacidade invasiva do *Streptococcus pneumoniae*. Os genes associados ao potencial invasivo são, assim, aqueles cuja utilização num modelo de classificação conduzem a *VPI* mais elevados.

Os resultados deste estudo revelam que é possível, a partir do conhecimento da presença ou ausência de um número reduzido de genes, prever se uma estirpe é ou não invasiva com uma exatidão relativamente elevada (86% quando são obtidos os grupos através da distância euclideana e classificação é efetuada a partir da matriz indicadora de presença), embora para as outras classes isso não se verifique.

A dificuldade em classificar as estirpes neutras poder-se-á dever ao seu caráter intermédio (o *OR* do respetivo serotipo não permitiu tirar conclusões claras acerca do seu potencial invasivo). No que diz respeito às estirpes colonizadoras, uma explicação possível para os seus valores preditivos baixos relativamente aos *VPI* pode ser o facto de os genes necessários à capacidade colonizadora do *Streptococcus pneumoniae* se encontrarem por defeito em todas as estirpes, inclusive as mais invasivas.

A aquisição de novos conhecimentos acerca das estirpes do *Streptococcus pneumoniae* e do seu genoma, e o aprofundamento dos métodos desenvolvidos neste estudo poderão no futuro otimizar a classificação das estirpes através de modelos de classificação em árvore no sentido de melhorar o seu desempenho.

A

Códigos em R

1.1. Função para cálculo do Coeficiente de Concordância Simples

Calcula o coeficiente de concordância simples para as colunas da matriz dada; n11,n00,n10 e n01 são matrizes de concordâncias. Recebe como argumento uma matriz binária **x** e retorna uma matriz de distâncias **c**.

```
concsimples<-function(x){
        n11<-matrix(0,nrow=ncol(x),ncol=ncol(x))
        n00<-matrix(0,nrow=ncol(x),ncol=ncol(x))
        n01<-matrix(0,nrow=ncol(x),ncol=ncol(x))
        n10<-matrix(0,nrow=ncol(x),ncol=ncol(x))
        n<-ncol(x);n
                for(i in 1:(n-1)){
                        for(j in (i+1):n){
                n11[i,j]<-sum(x[,i]*x[,j])
                n11[j,i]<-n11[i,j]
                n00[i,j]<-sum((1-x[,i])*(1-x[,j]))
                n00[j,i]<-n00[i,j]
                n10[i,j]<-sum(x[,i]*(1-x[,j]))
                n10[j,i]<-n10[i,j]
                n01[i,j]<-sum((1-x[,i])*x[,j])
                n01[j,i]<-n01[i,j]
                        }
                }
                for(i in 1:n){
                n11[i,i]<-sum(x[,i]*x[,i])
                n00[i,i]<-sum((1-x[,i])*(1-x[,i]))
                }
        c<-matrix(0,n,n)
        c<-(n11+n00)/nrow(x)
}
```

1.2. Função para cálculo da entropia do sistema

Recebe um vetor de desfechos correspondentes a cada objeto (**x**)e retorna o valor da entropia do sistema, **H**. Neste caso existem três desfechos: C, I e N.

```
ent.sist<- function(x){
        n<-length(x)
        u<-c("C","I","N")
        a<-factor(x,levels=u)
        t<-table(a)
        ti<-t/sum(t)
        H<-0
        for(i in 1:length(ti)){
                if(ti[i]!=0)
                H<-H+(-ti[i]*log(ti[i],2))
        }
return (H)
}
```

1.3. Função para cálculo da entropia condicional a um conjunto de variáveis.

Recebe a matriz **g** das variáveis (dimensão p×n) e o vetor de desfechos **v** de dimensão n. Retorna uma matriz com três colunas, correspondentes à identificação de cada variável (gene), à entropia condicional à mesma (entropia) e ao ganho associado à sua inclusão no sistema (ganho), por ordem decrescente de ganho.

```
ent<-function(g,v){
        eCond<-matrix(0,nrow=nrow(g),ncol=2)
        for (jj in 1:nrow(g)){
                levels(v)<-c("C","I","N")
                gen<-factor(unlist(g[jj,]))
                levels(gen)<-c("0","1")
                t<-table(gen,v)
                tp<-prop.table(t,1) # probs condicionais
                if(sum(t[1,])==0|sum(t[2,])==0)eCond[jj,1]<-H
                else{
                        k<-dim(tp)[1]  # valores do gene (0 ou 1)
                        cy<-dim(tp)[2] #n° colunas=n° categorias de y: C,I,N
                        p<-rep(0,k)
                        p<-rowSums(t)/sum(t)
                        ent<-tp
                        for(i in 1:k)
                                for(j in 1:cy)
```

```
                                        if(ent[i,j]!=0)ent[i,j]<--
        ent[i,j]*log2(ent[i,j])
                        eCond[jj,1]<-sum(rowSums(ent)*p)
                }
                eCond[jj,2]<-H-eCond[jj,1]
        }
entropia<-cbind(b,eCond)
vranks<-nrow(g)-rank(entropia[,3])+1
ovranks<-order(vranks)
sortent<-entropia[ovranks,]
names(sortent)<-c("gene","entropia","ganho")
return(sortent)
}
```

1.3. Algoritmo da validação cruzada *leave one out* (2.3.3.1)

Função que calcula a classificação da estirpe em cada ciclo de validação cruzada. O resultado final é a última classificação. Recebe a matriz de variáveis **g**, o vetor de desfechos **v** e a identificação da variável que determina a partição **qual.**

```
class.est<-function(qual,g,v){
        gene<-factor(unlist(g[qual,]))
        levels(v)<-c("C","I","N")
        levels(gene)<-c("0","1")
        t<-table(gene,v)
                tp<-prop.table(t,1) # probs condicionais
                if(sum(t[1,])==0){
                        tp[1,]=0
                }
                if(sum(t[2,])==0){
                        tp[2,]=0
                }
        gTot<-read.table("ncore.txt",head=T)
        estado<-gTot[qual,i]
        class<-which(tp[estado+1,]==max(tp[estado+1,]))
        if(length(class)==1)escolha<-class
        if(length(class)==2){
                u<-runif(1,0,1)
                ifelse(u<0.5,escolha<-class[1],escolha<-class[2])}
      if(length(class)==3){
                u<-runif(1,0,1)
                escolha<-class[3]
                if(u<2/3)escolha<-class[2]
                if(u<1/3)escolha<-class[1]
```

```
        }
        parar<-FALSE
      if(max(tp[estado+1,]==1))parar=TRUE
        return(names(escolha))
}
```

Função que, a partir dos mesmos argumentos da função class.est, permite parar o processo de classificação da estirpe quando esta é classificada numa categoria com probabilidade igual a 1 (objetos do ramo todos pertencentes à mesma classe).

```
para<-function(qual,g,v){
        gene<-factor(unlist(g[qual,]))
        levels(v)<-c("C","I","N")
        levels(gene)<-c("0","1")
        t<-table(gene,v)
        tp<-prop.table(t,1) # probs condicionais
        if(sum(t[1,])==0){
                        tp[1,]=0
                }
                if(sum(t[2,])==0){
                        tp[2,]=0
                }
        gTot<-read.table("ncore.txt",head=T)
        estado<-gTot[qual,i]
        parar<-FALSE
      if(max(tp[estado+1,]==1))parar=TRUE
        return(parar)
}
```

Código para a classificação das estirpes, a partir do qual se obtêm as variáveis que presidem a cada partição (classCiclo), a classificação das mesmas em cada uma das partições (categCiclo) e a classificação final de cada uma das estirpes (classFinal), a partir da qual são calculadas a precisão e valores preditivos.

```
inv<-read.table("inv.txt",head=T)
v<-inv[,2]
classe<-v
classe.est.passos<-matrix(999,nrow=20,ncol=72)
classif.est<-rep(999,72)
g<-matrix(0,nrow=20,ncol=72)
for (i in 1:72){
      gen<-read.table("ncore.txt",head=T)
      inv<-read.table("inv.txt",head=T)
      v<-inv[,2]
      gen<-gen[,-i]
```

```
        v<-v[-i]
       H<-rep(999,20)
        j<-0
        parar<-FALSE
        while(parar==FALSE & j<20){
                j<-j+1
                H[j]<-ent.sist(v)
                sortent<-ent(gen,v,H[j])
                qual<-as.numeric(rownames(sortent[1,]))
                g[j,i]<-sortent[1,1]
                classe.est.passos[j,i]<-class.est(qual,gen,v)
                classif.est[i]<-classe.est.passos[j,i]
                gTot<-read.table("ncore.txt",head=T)
                estado<-gTot[qual,i]
                parar=para(qual,gen,v)
                iguais<-which(gen[qual,]==estado)
                gen<-gen[,iguais]
                v<-v[iguais]
        }
}
write.table(g[1:8,1:72],"classCiclo.txt")
classe.est.passos[,1:72]
write.table(classe.est.passos[1:8,1:72],"categCiclo.txt")
classif.est[1:72]
write.table(classif.est[1:72],"classFinal.txt")
```

1.4. Algoritmo de seleção de genes para elaboração do modelo em árvore

Efetua a classificação hierárquica dos genes a partir de uma matriz de distâncias em formato *dist* (as.dist), divide o dendograma resultante em grupos de 2 a um número máximo pré-definido (ngr), seleciona o gene que, em cada grupo, tem o valor máximo de ganho (a partir dos resultados da função ent, guardados no objeto **entropia**), e coloca a identificação dos genes representativos dos grupos na matriz designada por **m**, que irá ser argumento das funções crossVal e crossValCP.

Exemplo da aplicação para a matriz de distâncias **dgsimples**:

```
dend<-hclust(dgsimples,method="complete",members=NULL)
plot(dend, main="Genes (Ligação Completa)",cex=0.7)
m<-matrix(0,ngr,ngr-1)
for(j in 2:ngr){
        grupo<-rect.hclust(dend, k=j, border="red")
        max<-rep(0,j)
        maxent<-matrix(0,nrow=ngr,ncol=j)
```

```
        for (k in 1:length(grupo)){
                h<-grupo[[k]]
                z<-rep(0,length(h))
                        for(i in 1:length(h)){
                        z[i]<-which(entropia[,1]==names(h[i]))
                        }
        parc<-entropia[z,]
        x<-max(parc[,3])
        y<-which(parc[,3]==x)
        max[k]<-parc[y,1]
        maxent[,k]<-g[,which(colnames(g)==max[k])]
        }
m[1:j,j-1]<-max
}
```

1.5. Algoritmo de classificação com validação cruzada *leave one out*, utilizando as funções *rpart* e *rpart.control* (*package* Rpart)

Função que recebe como argumentos a matriz **m** dos grupos, a matriz das variáveis **g** (binária ou de probabilidades) e o vetor de desfechos **iv**. Retorna a precisão dos classificadores e os valores preditivos.

```
crossVal<-function(m,g,iv){
K=72
res<-matrix(0,ncol(m),5)
colnames(res)<-c("err.cv","acc","vpc","vpi","vpn")
for(i in 1:ncol(m)){
        a<-m[1:(i+1),i]
        gn<-matrix(0,nrow=length(iv),ncol=(i+1))
                for(j in 1:length(a)){
                        gn[,j]<-g[,colnames(g)==a[j]]
                }
        colnames(gn)<-a
        gn<-cbind(iv,gn) #gn passa a ser a matriz de dados
        rownames(gn)<-rownames(g)
        gn<-as.data.frame(gn)
        write.table(gn,"gprovisorio.txt")
        gn<-read.table("gprovisorio.txt",head=T)
        predT<-rep(0,K)#onde vão ser colocadas as classificações à posteriori
        classerr<-rep(0,K)
        iv<-gn$iv
        for(k in 1:K){
                cont<-rpart.control(minbucket=1)
                arVal<-rpart(iv ~. ,data=gn[-k,],
method="class",parms=list(split="information"),control=cont)
```

```
                pred=predict(arVal,newdata=gn[k,],type="class",
parms=list(split="information"))
                ifelse(iv[k]==pred,classerr[k]<-0,classerr[k]<-1)
                predT[k]<-pred
        }
        #Erro da estimação da validação cruzada:
        res[i,1]<-sum(classerr)/K #erro da valid. cruzada err.cv
        res[i,2]=1-res[i,1] #precisão
        mc<-table(predT,iv)
        res[i,3]<-mc[1,1]/sum(mc[1,]) #vpc
        res[i,4]<-mc[2,2]/sum(mc[2,]) #vpi
        res[i,5]<-mc[3,3]/sum(mc[3,]) #vpn
}
return(res)
}
```

1.6. Algoritmo de classificação com validação cruzada *leave one out* a partir das componentes principais das variáveis, utilizando as funções *rpart, rpart.control* e *princomp*

Recebe os mesmos argumentos da função crossVal, e retorna uma matriz com a mesma informação, mas a classificação é efetuada a partir das componentes principais das variáveis de entrada, calculada através da função **princomp** do R.

```
crossValCP<-function(m,g,iv){
K=72
res<-matrix(0,ncol(m),5)
colnames(res)<-c("err.cv","acc","vpc","vpi","vpn")
for(i in 1:ncol(m)){
        a<-m[1:(i+1),i]
        gn<-matrix(0,nrow=length(iv),ncol=(i+1))
                for(j in 1:length(a)){
                        gn[,j]<-g[,colnames(g)==a[j]]
                }
        colnames(gn)<-a
        gcp<-princomp(gn)#ACP com o grupo de genes 2 a 72
        gnacp<-gcp$scores #Matriz de scores<-matriz de dados
        gnacp<-cbind(iv,gnacp)
        rownames(gnacp)<-rownames(g)
        gnacp<-as.data.frame(gnacp)
        write.table(gnacp,"gacprovisorio.txt")
        gn<-read.table("gacprovisorio.txt",head=T)
        predT<-rep(0,K)#onde vão ser colocadas as classificações à posteriori
```

```
        classerr<-rep(0,K)
        iv<-gnacp$iv
        for(k in 1:K){
                cont<-rpart.control(minbucket=1)
                arVal<-rpart(iv ~. ,data=gnacp[-
k,],method="class",parms=list(split="information"),control=cont)
                pred=predict(arVal,newdata=gnacp[k,],type="class",
parms=list(split="information"))
                ifelse(iv[k]==pred,classerr[k]<-0,classerr[k]<-1)
                predT[k]<-pred
        }
        #Erro da estimação da validação cruzada:
        res[i,1]<-sum(classerr)/K #erro da valid. cruzada err.cv
        res[i,2]=1-res[i,1] #precisão
        mc<-table(predT,iv)
        res[i,3]<-mc[1,1]/sum(mc[1,]) #vpc
        res[i,4]<-mc[2,2]/sum(mc[2,]) #vpi
        res[i,5]<-mc[3,3]/sum(mc[3,]) #vpn
}
return (res)
}
```

B

Tabela de classificação das estirpes

Nome da estirpe	*Classificação*
X1999V0053S	N
X1999V0906S	I
X1999V0980S	I
X1999V0993S	N
X1999V1040S	I
X1999V1216S	I
X2000V0189S	C
X2000V0324S	N
X2000V0527S	I
X2000V0626S	I
X2000V0637S	I
X2000V0731S	C
X2000V0734S	C
X2000V0926S	C
X2000V1024S	I
X2000V1277S	C
X2001V0050S	C
X2001V0240S	I
X2001V0381S	C
X2001V0406S	N
X2001V0596S	I
X2001V0617S	C
X2001V0959S	C
X2002V0068S	I
X2002V0086S	C
X2002V0307S	N
X2002V0321S	I
X2002V0361S	N
X2002V0380S	C
X2002V0615S	N
X2002V0701S	N
X2003V0336S	N

Nome da estirpe	***Classificação***
X2003V0590S	C
X2003V0851S	C
X2003V1045S	N
X2003V1049S	I
X2003V1114S	N
X2003V1198S	N
X2003V1366S	I
X2003V1397S	I
X2003V1491S	N
X2003V1526S	N
X2003V1580S	C
X2004V0441S	N
X2004V1384S	I
X2005V0492S	N
X2005V0786S	I
X2005V1379S	C
X2005V2089S	N
X2003V1425S	I
X2005V0143S	I
X2003V0108S	C
X2001V0619S	N
PP71S	C
X2003V0342S	I
X2003V1050S	I
X2003V0173S	I
X2004V0126S	N
X2005V1197S	N
X2003V0902S	C
PP23S	C
X1999V0341S	I
X2003V1593S	I
X2003V0364S	I
X2005V0806S	C
X2001V0098S	I
X2000V1177S	I
X2001V1092S	I
X1999V1095S	I

Nome da estirpe	*Classificação*
X2003V0954S	C
X2003V0286S	N
X1999V1076S	N

Informação respeitante a alguns genes utilizados na classificação supervisionada

Gene	*Oligo.ID*	*Primary Target*	*Gene Symbol*	*Common name of primary target*
104	7QSP00002_F_11	SP0142		hypothetical protein
185	7QSP00008_O_24	SP0250		PTS system, IIC component
754	7QSP00010_N_3	SP1017	xylH	4-oxalocrotonate tautomerase
972	7QSP00009_A_8	SP1317	ntpA	v-type sodium ATP synthase, subunit
1058	7QSP00004_E_9	SP1437		conserved domain protein
1181	7QSP00003_P_8	SP1620		PTS system, nitrogen regulatory component IIA, putative
1223	7QSP00009_N_3	SP1680		conserved hypothetical protein
1619	7QSP00009_F_15	SPN01012		SURFACE PROTEIN C PSPC (FRAGMENT).
2118	7QSP00009_N_3	SPN05013		ORF16.
2453	7QSP00004_B_9	SPN07050		HYPOTHETICAL PROTEIN 170 AACS LONG
2593	7QSP00005_J_5	SPN08232		25.5 KDA PROTEIN (PUTATIVE CHAIN LENGTH REGULATOR)
2605	7QSP00006_A_1	SPN08245		DTDP-L-RHAMNOSE SYNTHASE
2777	7QSP00008_I_24	SPN12029		HYPOTHETICAL PROTEIN 430 AACS LONG

B

Gene	*Oligo.ID*	*Primary Target*	*Gene Symbol*	*Common name of primary target*
2971	7QSP00003_C_17	SPN23005		HYPOTHETICAL PROTEIN.
3055	7QSP00002_F_5	spr0105	Transporter -truncation	Transporter, truncation
3186	7QSP00009_P_21	spr0492		valyl-tRNA synthetase
3193	7QSP00002_A_22	spr0501	valS	hypothetical protein
3393	7QSP00005_G_6	spr1183	ABC-NBD-truncation	ABC transporter ATP-bindng protein - possibly multidrug efflux, truncation
3522	7QSP00006_K_21	spr1631	trpA	tryptophan synthase subunit alpha

5. Bibliografia

Antunes M. and Sousa L. **Bayesian Classification and Non-Bayesian Label Estimation via EM Algorithm to Identify Differentially Expressed Genes: a Comparative Study**. *Biometrical Journal* 50,5, 824-836. (2008).

Antunes M. **CRM e Prospecção de Dados** - Textos de apoio à disciplina de CRM e Prospecção de Dados, Centro de Estatística e Aplicações da Universidade de Lisboa (2010).

Breiman L., Friedman J. H., Stone C.J., Olshen, R.A. **Classification and Regression Trees.** *Chapman & Hall CRC* (1984).

Bogaert D., Groot R., and Hermans, P. W. M. ***Streptococcus pneumoniae*** **colonisation: the key to pneumococcal disease.** *THE LANCET Infectious Diseases* Vol 4 March (2004).

Bröet, P., Richardson, S., and Radvanyi, F. **Bayesian hierarchical model for identifying changes in gene expression from microarray experiments.** *Journal of Computational Biology* 9, 671–683 (2002).

Calix, J. J. and M. H. Nahm. 2010. **A new pneumococcal serotype, 11E, has a variably inactivated wcjE gene**. *J. Infect. Dis.* 202**:**29–38 (2010).

Cardoso, L. **Classificação de genes em hibridação genómica comparativa de estirpes de *Streptococcus pneumoniae***. Dissertação de Mestrado em Bioestatística na Faculdade de Ciências da Universidade de Lisboa (2009).

Claverys J., Prudhomme M., Mortier-Barrière I., Martin B. **Adaptation to the environment: Streptococcus pneumoniae, a paradigm for recombination-mediated genetic plasticity?** *Molecular Microbiology*, 35(2), 251-259 (2000).

Dean N. and Raftery A.E**. Normal uniform mixture differential gene expression detection for cDNA microarrays**. *Genome Biology* 6:173 (2005).

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. **Empirical Bayes Analysis of a Microarray Experiment.** *Journal of the American Statistical Association* 96, 456, 1151–1160 (2001)

Bramer M. **Principles of Data Mining.** *Springer* (2007).

Griffiths A., Miller J., Suzuki D., Lewontin R., Gelbart W., **An Introduction to Genetic Analysis,** 6ª Edição - *Freeman* (1996).

Hand D., Mannila H., Smyth P. **Principles of Data Mining.** *The MIT Press* (2001).

Hastie T., Tibshirani R., and Friedman J. **The elements of statistical learning: data mining, inference, and prediction**, 2ªEdição- *Springer Series in Statistics.* (2009).

Hiller N.L., Ahmed A., Powell E., Martin D. P., Eutsey R., Earl J., Janto B., Boissy R.J., Hogg J., Barbadora K., Sampath R., Lonergan S, Post J. C.,6,7, Hu F. Z.,Ehrlich G. D. **Generation of Genic Diversity among Streptococcus pneumoniae Strains via Horizontal Gene Transfer duringa Chronic Polyclonal Pediatric Infection.** *PLoS Pathogens* V. 6, Issue 9,e1001108 (2010).

Jiang L., Li C., **An Empirical Study on Attribute Selection Measures in Decision Tree Learning** *Journal of Computational Information Systems* 6:1;105-112 (2010).

Kendziorski, C. M., Newton, M. A., Lan, H. and Gould, M. N. **On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.** *Statistics in Medicine* 22, 3899–3914 (2003)

Leal, M. - Diapositivos das aulas da disciplina de **Análise de Dados Multivariados** - mestrado em Bioestatística (ano letivo 2009/2010).

Lee, M. L. T., Lu, W., Whitmore, G. A., and Beier, D. **Models for microarray gene expression data.** *Journal of Biopharmaceutical Statistics* 12, 1–19 (2002).

Lönnstedt, I. and Speed, T. **Replicated Microarray Data.** *Statistica Sinica* 12, 31–46 (2002).

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. **On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray** *Data. Journal of Computational Biology* 8(1), 37–52 (2001).

Obert C., Sublett J., Kaushal D., Hinojosa E., Barton T., Tuomanen E. I., Orihuela C. J. **Identification of a Candidate *Streptococcus pneumoniae* Core Genome and Regions of Diversity Correlated with Invasive Pneumococcal Disease**. *Infection and Immunity*, Vol.74 August, No. 8, p. 4766-4777.

Pan, W. **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments**. *Bioinformatics* 18, 546–554 (2002).

Pinto F. R., Aguiar S. I., Melo-Cristino J. and Ramirez M. **Optimal control and analysis of two-color genomotyping experiments using bacterial multistrain arrays.** *BMC Genomics* 9:230 (2008).

Rebouças, S. **Metodologias de Classificação Supervisionada para Análise de Dados de *microarrays*** Tese de Doutoramento em Estatística e Investigação Operacional na Faculdade de Ciências da Universidade de Lisboa, especialidade de Probabilidades e Estatística (2011).

Sá-Leão R., Pinto F., Aguiar S., Nunes S. Carriço J., Frazão N., Gonçalves-Sousa N., Melo-Cristino J., Lencastre H., Ramirez M. **Analysis of Pneumoccoccal Serotypes and Clones Circulating in Portugal before Widespread Use of Conjugate Vaccines Reveals Heterogenous Behaviour os Clones Expressing the Same Serotype.** *Journal of Clinical Microbiology*, Vol. 49, No 4 (2011).

Snipen L., Repsilber D., Nyquist L., Ziegler A., Aakra A. and Aastveit A. **Detection of divergent genes in microbial aCGH experiments**. *BMC Bioinformatics* 7:181. (2006).

Draghici, S. **Data analysis for DNA microarrays** Chapman & Hall/CRC

Storey, J. D. and Tibshirani, R. **Statistical significance for genome-wide studies.** *Proceedings of the National Academy of Sciences* 100, 9440–9445 (2003).

The R project for Statistical Computing. http://www.r-project.org/.